

BIG DATA DIGEST 大数据文摘

战略数据合作方 Linkedin 领英

顶级数据团队建设全景报告

A ROADMAP TO A TOP DATA-DRIVEN ENTERPRISE

A ROADMAP TO
A TOP DATA-DRIVEN ENTERPRISE

2020

VISIT…

LANZAROTE Caliente.COM

扫码免费领取资料



房中术(www.zuihaoziyuan.com)专注于男性增大增粗增长、阳痿、早泄、壮阳、延时、强肾、回春、健身。女性缩阴、丰胸、减肥、化妆、瑜伽、保养、产后修复、盆底肌锻炼。两性健康,夫妻按摩,房中术,性姿势,性技巧,性知识等

更多免费教程：英语学习，技能提升，PS 教学，投资赚钱，音乐教程，口才教学，情商提升，风水教学，心理学，摄影知识，幼儿学习，书法学习，记忆力提升等等......

扫码加微信lax988988领取资料



==

全站课程下载

课程不断增加

本站现资源容量已超 10T

入群联系 QQ：167520299 或添加微信：1131084518（备注PDF）

阳痿早泄训练
皇室洗髓功视频教学
女人驻颜术
泡妞约炮万元课程
足疗养 SPA 教材
玉蛋功
马氏回春功
房中术张丰川
哲龙全套视频
增大盼你增大
国际男优训练
亚当德永早泄训练
洗髓功真人内部
皇室养生绝学道家洗髓功
【铁牛人会员课】男人必备技能，理论讲解
实战高清视频
随意控制射精锻炼 视频+图片+文字
价值 1440 元第一性学名著<素女经房中养生
宝典视频>12 部
洗髓功修炼方法视频教学
陈见玉蛋功视频教学 女性缩阴锻炼
男性自然增大增长指南
强性健肾保健操 1-4
道家强肾系统锻炼功法
马氏回春功
12 堂课，全面掌握男性健康问题 让你重燃自信

联系微信：1131084518

1、东方性经
2、印度 17 式
3、口交技巧 3 部
4、港台性姿势 3 部
5、365 性姿势 6 部
6、泰国性爱密经 17 式
7、花花公子性技巧 6 部
8、阁楼艳星性技巧 7 部
9、古今鸳鸯秘谱全集 7 部
10、夫妻爱侣情趣瑜伽 2 部
11、古代宫廷性保健系列 14 部
12、汉唐宋元明清春宫图真人
13. 柔软性爱宝典 日本 9800 课
14. 李熙墨 3999 全套课
15. 妖精性爱课 2888
16. 李银河全套性课
17. 领统统性课
18. 德勇男性篇
19. 德勇男性篇
20. 缓慢性爱
21. 亚当多体位搭配篇
22. 亚当多体位结合篇
23. 德勇克服早泄讲座练习
24. 德勇以女性为中心得爱抚
25. 加藤鹰接吻爱抚舌技
26. 加藤鹰指技
27. 加藤鹰四十八手入门
28. 佐藤潮吹教学
29. 佐藤男人体能锻炼+保健品介绍
30. 佐藤男人早泄对抗训练
31. 阿拉伯延时训练
32. 田渊正浩秘籍
33. 异性性快感集中训练教学
34. 自我愉悦锻炼密宗
35. 铁牛全套延时训练课
36. Pc 机锻炼真人视频教学
37. 印度性经全集 8 部
38. 21 世纪性爱指南
39. 香蕉大叔男女训练馆全套
40. 中美真人性治疗教学+理论
41. 女性闺房秘术
42. 幸福玛利亚性课
43. 陈见如何释放性魅力征服
44. 性爱技巧讲座全套
45. 性爱秘籍全套
46. 性爱误区讲座
47. 性病讲解大全
48. 性博士讲座合集
49. 性健康和性高潮合集
50. 性教育讲座合集
51. 性能力课堂合集
52. 性生活问题解析合集
53. 意外怀孕和避孕处理课堂
54. 性感地带探索
55. 性技巧讲座
56. 性健康与性卫生讲座
57. 性生活专家答疑
58. 性心理与性道德合集
59. 性爱宝典合集
60. 性爱技巧合集
61. 完美性爱演示
62. 完美性爱技术讲解

更多精品等你来解锁哦............

每日免费获取报告

1、每日微信群内分享7+最新重磅报告；

2、每日分享当日华尔街日报、金融时报；

3、每周分享经济学人

4、行研报告均为公开版，权利归原作者所有，起点财经仅分发做内部学习。



扫一扫二维码
关注公号
回复：研究报告
加入“起点财经”微信群。。

目录

数据行业概况 03

1.1 数据行业概况 04
1.2 大数据行业发展及团队建设困境 04
1.3 数据团队组建运营方法论 05

数据团队宏观发展现状 08

2.1 职位概况 09
2.2 工作经验 10
2.3 学历要求 11
2.4 薪资水平 12

数据团队从业者微观洞察 13

3.1 工具、技术、方法论 14
3.2 自我认知 16
3.3 工作生活 16
3.4 数据应用场景 16
3.5 团队前景 17
3.6 大数据与人工智能问题 17

全球数据团队观察 19

4.1 学历分布和毕业院校 20
4.2 主流专业分布 21
4.3 数据人才集中分布的行业 21
4.4 数据人才的跨行业流动 22
4.5 数据人才的主流职位和增长最迅速的数据相关职位 23
4.6 数据从业者集中分布的地区 24

疫情中的数据行业 26

附录 29

BIG DATA DIGEST 大数据文摘 清华大学软件学院



报告发布方





BIG DATA DIGEST
大数据文摘

战略数据合作方

LinkedIn领英

调研方法

联合调研组采用了海量数据分析、定向问卷调查与深度访谈等方法，分别针对企业高层、数据团队负责人、数据从业者和其他相关人员进行广泛而深入的调研，力求从尽量多的角度还原现阶段数据团队的建设全景。

- 海量数据分析：对“数据”、“分析”、“机器学习”等关键词进行全网爬取，通过数据清洗、数据分析、数据可视化等步骤对 12 万条网络公开招聘信息进行分析。

- 定向问卷调查：通过互联网向数据团队相关从业者和负责人发放定向问卷，并回收近千份有效问卷。

- 深度访谈：对 6 位优秀数据团队负责人进行深度访谈，涵盖国内外不同行业及发展阶段的公司或组织。



PART 1

数据行业概况

BIG DATA DIGEST 大数据文摘
清华大学软件学院

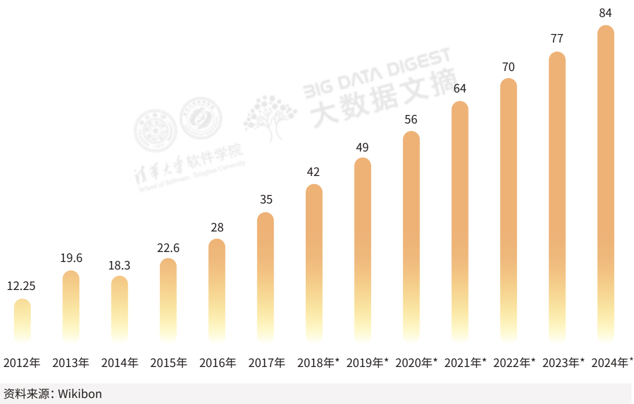
1.1 数据行业概况

大数据一词对于大众来说，已经不再陌生。根据 Wikibon 研究数据，作为大数据产业发展的基石，全球的大数据市场规模预计从 2018 年的 420 亿美元增长至 2024 年的 840 亿美元。（如图 1-1-1）

在中国，自 2015 年国务院颁布《促进大数据发展行动纲要》后，大数据正式上升为国家发展战略，而 2016 年由工信部印发的《大数据产业发展规划（2016-2020 年）》则掀起了大数据产业建设的浪潮。根据赛迪数据与《2019 年中国大数据行业研究报告》，2018 年中国大数据产业规模达到 4384.5 亿元，同比增长 23.5%，据预测到 2021 年，中国大数据产业规模将超过 8000 亿元。(如图 1-1-2)

在具体的大数据应用领域，生态环境、农业、水利、医疗、交通旅游服务等都在一系列政府政策的支持下，成为了大数据技术的实际应用场景；而政府资源的支持与技术实力的稳步增强，为这些领域的大数据实践打下了至关重要的基础。

图1-1-1 2012-2024 年全球大数据市场规模（单位：十亿美元）

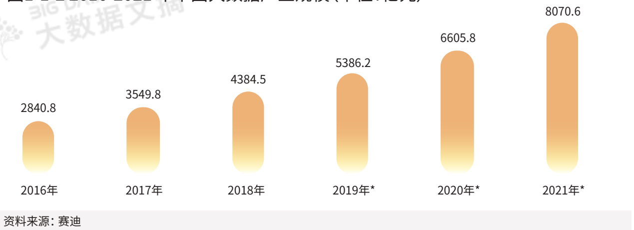


资料来源：Wikibon

1.2 大数据行业发展及团队建设困境

数据团队是近些年随着大数据概念的推广而产生的新型团队。因此，相较于组织或机构内成熟运行的其他部门，数据团队的成立时间较短。在政策扶持与业界高度关注下，这些团队的发展和建设过程中往往面临一些问题。

图1-1-2 2016-2021 年中国大数据产业规模（单位：亿元）



资料来源：赛迪

一、专业人才缺口大，行业间分布不平均

专业的数据人才缺乏，这是整个数据行业面临的第一大困境。经过最近几年高校的大数据专业设置以及相关大数据职业培训，中低层数据分析人才逐渐成熟并进入市场，满足了一些数据团队基础职位需求，但是，顶级和有丰富经验的数据人才依然难求。

调研过程中我们发现，互联网行业对数据人才的吸引力最高，规模大、薪水高的大公司对于人才的吸引力也一直居高不下。相对而言，初创企业和刚刚进行数字化转型的传统企业往往难觅经验丰富的数据团队领头人。此外，高等教育领域的数据人才流出现象严重，高校和研究院“留不住人”的现象在最近几年愈演愈烈。

二、高层管理者支持力度不够，建设目标不清晰

目前，行业内已经达成了普遍共识：数字化转型和数据团队建设是“一把手”工程，需要领导层对数据驱动保有完整的认知和贯彻决心，以完成“自上而下”的推动。

很多企业中，高层管理者自己并没有想好是否要进行数字化转型，就“赶时髦”开始了建设，遇到困难后立刻回到了固有的传统管理模式，这些企业的数据建设目标不清晰，因此很难获得成效。

三、企业数字驱动文化建设不完善，业务界限模糊

除了高层的认知，整个公司的数据文化也对于数据团队的建设至关重要。这又与企业本身的文化和属性密不可分。例如工程师文化浓厚、创新意识强的互联网公司往往在数据团队建设和数字化转型中进展更顺利、快速，而不够开放务实的企业往往在数字化转型中推进困难。

1.3 数据团队组建运营方法论

失败的团队各有各的问题，成功的团队却一定是相似的。

不同的行业、企业的数据团队，虽然建设路径有所不同，但从整体来看，“成功”的数据团队都有着相似的发展阶段、趋势，和一套明确的价值衡量体系。

1.3.1 数据团队的发展阶段

数据团队的组建运营在不同行业、公司各有不同，但一般会随着企业的数字化，经过以下三个发展阶段，数据团队在每个阶段需要配合公司发展的相应能力。

第一阶段：大数据基础平台建设

还没有进行数字化转型的企业，数据往往分散、杂乱无章地散落在各事业部内部，形成孤岛。

数据团队建设的第一个阶段，需要首先对数据进行集成、清洗和集合，或者说数据治理。在这个阶段，数据团队需要先完成大数据的一些基础设施建设，同时围绕自身业务模式，建设数据平台，比如用户数据平台等。

需要注意的是，从数据资产的梳理、清洗加工、结构化分析、产出相应形式的数据产品，到最后的投入使用，其生产链条之长，涉及的角色和变量之多，都为数据团队的工作提出了极高要求。同时，伴随着业务系统的变化，对应的数据逻辑，以及指标口径的定义也会随之变化。需要在这个过程中密切围绕用户和核心业务进行调研，尽量平衡效率和成本。

第二阶段：数据分析产出提升核心业务

在完成了第一阶段的大数据基础平台建设和基本工作流程建设后，数据团队开始产出大量的数据分析结果，一般包括行业趋势分析、市场销量和价格预测、战略和竞品策略分析等，辅助业务团队和管理团队进行数据驱动的决策。

同时这一阶段，数据团队会开始建设可扩展的数据分析解决方案：在运营实施过程中进行流程优化和自动化，比如对于多个业务团队需求、复用性高的项目，可以通过A/B测试，迅速建立流程化的操作平台，业务团队成员和工程师可以直接在平台上创建调试，产品经理也可以在测试开始几小时后直接在平台上查看测试结果，最大化的减少数据科学家的重复劳动。

这一工作流程的建立，对于一个数据团队和整个企业的发展意义重大，可以将数据科学家和分析师从繁杂的劳动中“拯救”出来，专注更有创造性的工作。

> 滴滴的数据科学团队每周会产出几千次的实验和评估，这些主要针对业务或产品的方案进行评估，更加自动化和流程化，随时可能影响到公司决策。
> ——滴滴技术副总裁、数据科学与智能部的负责人赖春波

第三阶段：数据文化和生态建设，包括数据驱动文化形成和人才培养

经过前两个阶段的建设，数据团队已经在公司内部建立起了一套比较完整的数据驱动平台，可以存放、调用核心数据和公用数据；业务部门使用者能够自行在平台上完成分析工作。

同时，为了保证整个数据团队的活力，公司内部的数据文化建设和人才培养机制也必不可少。例如给优秀的业务团队分析师设置系统的数据技能培训，能够让业务人员参与进数据驱动过程，将业务经验与数据分析方法更好地结合起来，以自服务式培养机制解决数据科学家缺失的难题。

在这一阶段，也有一些数据团队会开始展开行业生态建设，将数据能力输出到上下游企业：在公司内部先部署与使用，比较成熟之后，再服务外部客户，同时通过反馈去优化产品，增加内部应用平台的功能特性，并建立内外部互动模式。

> 联想希望在对外服务的同时，也能够建立起生态系统，服务更多的尤其是中小企业客户，推动中国智能制造快速转型。
> ——联想数据智能业务集团产品及生态总经理田日辉

1.3.2 数据团队发展趋势

近年来，随着大数据和人工智能的发展，数据科学领域囊括的范围越来越广泛。在发展中，数据团队正逐渐呈现“嵌入化”、“专业化”和“不唯数据论”的三大趋势。

一、“嵌入化”趋势：向业务团队靠拢

数据对于每家公司的重要性都只增不减，数据科学向业务方向的“嵌入”趋势越来越明显、边界也越来越模糊。也正因如此，数据团队的业务范围也逐渐扩大化。

除了比较常见与数据团队密切接触的市场、产品团队，一些偏底层和基础架构的项目也在积极寻求数据团队的支持，比如数据中心和工程团队，希望数据科学团队能够提供数据支持，减少算力能耗。

> 滴滴的数据分析师和数据科学家，以“嵌入式”的方式，分布在不同的业务部门中。数据科学团队，需要在业务形态中实现广泛的运营智能、产品智能和决策智能，助力业务可持续发展。
> ——滴滴技术副总裁、数据科学与智能部的负责人赖春波

> 走在业务线前面主动去做一些工作，每当业务碰到的问题时，最好平台都有解。
> ——美团数据平台负责人李闻

二、“专业化”趋势：基础设施建设与数据科学应用团队逐渐分离

在几年前，数据团队会囊括Hadoop、Kafka等底层架构工程师，但是，随着数据科学的应用范围逐渐扩大，现在这些大数据底层技术都变成了偏基础设施的内容，在狭义概念上，已经不再属于数据科学团队的范畴。

> 对于领英来说，数据科学团队的整体趋势更加走向专业化，他们的职责不再是建立数据基础设施或平台，而是怎样去使用数据科学和工程来最大化数据的价值。
> ——领英全球数据团队负责人许亚

三、“不唯数据论”：把握好数据的度

目前，数据化决策已经成为了行业共识，企业已经不再对“数据驱动”本身产生质疑，数据团队的工作难点在于，当评估业务中一个不可量化的任务时，如何把握好“度”。

业务发展往往是一个系统工程，很多环节面临的问题并没有一个完美的模型可以解决。一边需要具体技术团队不断搜集正确的数据、优化算法，另一边则需要企业更多考虑现实和线下场景的复杂性。在数据和业务经验产生冲突时，数据团队更需要思考如何更好地引入业务经验，与数据结果结合，产生更好的决策指标。

> 要依赖数据做决策，但不能只依赖数据做决策。
> ——滴滴数据治理和数据平台部负责人王勇

四、更加重视数据安全：从被动应对到主动出击

产业互联网的发展，尤其是随着年初疫情的爆发，加快了实体业务数字化上云的步伐，随之而来的，企业的安全防护特别是数据安全的防护也在面临新的挑战。数据安全被不少数据团队提上日程，成为体现企业责任感的重要工作。

调研过程中我们发现，数据团队对安全的重视程度逐年增长，不少企业都专门设立了数据安全团队，从流程和企业文化层面加强了安全建设。在高速发展的产业互联网时代，大量新技术的出现让安全问题已经不是单点领域问题，而是一个系统工程，数据安全的应用范围早已超出了数据本身的安全，还涉及到整个安全体系。

> 将安全意识下沉到公司内部一定是自上而下的过程。近几年安全方面的立法立规不断增强，从欧洲 GDPR 到美国 CCPA，中国也出台了一系列数据安全法规。这其实就是从顶层驱动告诉大家，安全的重要性。
> ——腾讯安全副总裁黎巍

1.3.3 数据团队内部价值衡量与商业 KPI 的设定

数据团队相对复杂的构成、再加上和业务团队合作的紧密性，这些决定了量化数据团队的商业影响力和设置其发展路线不是一件容易的事。

不同数据团队面临不同的商业考量，虽然数据团队的价值很难量化，但仍然有些指标可以作为探讨的基础。访谈过程中，研究组也总结了数据团队内部考核自身商业影响力的几个重要因素。

一、数据易得性

数据易得性，即当外界需要数据时，获取相应数据的难易程度。在公司内部，数据团队拥有许多数据资源，比如原始数据，指标数据，数据模型，数据可视化。当外界对这些资源有需要的时候，如何能够保证这些需求能够随时被满足？

软件开发有一系列衡量数据获取难易程度的指标，比如 SLA(Service-Level Agreement) 的达标率就是一个很好的量化指标。

二、工作效率

工作效率，即个人的工作成果是否可以提升整个团队的工作效率。

数据科学经常面临这样的问题，做完一个有价值的分析后，并不关心后续的自动化过程，所以每次出现新需求，就需要再手动跑模型，造成了不同的人在做相同的重复劳动的局面。如果分析过程实现自动化，重复劳动时间被解放，整个数据科学团队的集体工作效率就都提高了。

> 数据科学家之前被人诟病过于追求新鲜感，喜欢挑战高难度问题，但做完 MVP (Minimum Viable Product) 后没有维护迭代的习惯，永远都在追逐下一个新难题。
> ——领英全球数据科学团队负责人许亚

三、战略化思维与直接商业影响力

战略化思维，即数据分析结果对公司重要战略性决策是否有指导作用，特别是工作成果对公司商业目标的直接影响力。

每个部门的工作开展都与公司的整体目标息息相关。数据团队在计划工作时，自然也需要考虑其目标能如何对公司的商业目标产生积极影响。

调研过程中我们发现，领英、滴滴、美团、联想等大型互联网企业的数据团队都会直接与公司高层建立沟通渠道，数据团队的产出也会直接影响公司的重要决策，因此，数据结果的准确性和前瞻性就至关重要。

例如领英需要了解用户在疫情期间是如何使用领英服务，如何通过领英的产品获取价值的，再就此对其数据团队的相应数据产品提出更实际的应用场景要求。

联想集团的数据团队也是通过直接的商业影响力衡量团队的价值。比如，首先通过全集团统一的数据平台了解公司内外对于数据平台的调用量级并设置 KPI；再通过相关操作了解平台数据调用为业务部门创造了哪些实际价值。

总体来说，数据团队的关键 KPI 包含用户数、直接产生的价值量等。由于最终的应用落地是由业务部门完成，项目完成后，业务效能提升如何，预测精准度表现如何，都对应着实际业务价值的产生，也是更容易衡量的指标。

PART 2

数据团队宏观发展现状

清华大学软件学院 BIG DATA DIGEST 大数据文摘

基于目前互联网的各大招聘网站公开信息，我们对于全国数据行业相关职位的整体情况进行了搜集，共采集 12 万份在招职位数据。基于这一数据集，我们将在本章探讨当前国内数据团队宏观发展现状。

2.1 职位概况

运营、数据、产品类需求旺盛

在全国数据行业相关职位中，运营、数据、产品、开发、网络类在招职位数量位列前五名，其中运营类占比 47.5%，数据类占比 27.3%，产品类占比约 10.8%。（如图 2-1-1）

一线、新一线城市梯队继续保持，深圳需求结构综合，北京侧重技术

所有数据类在招岗位中，深圳、广州、上海、北京、杭州五个城市的在招职位数量最多，随后则是成都、武汉、南京、长沙、苏州，呈现一线城市 - 新一线城市的梯队特征。下文在城市对比中，将重点观察深圳、广州、上海、北京、杭州五座城市的数据团队构成现状。（如图 2-1-2）

其中，观察每个城市在招岗位的类型特征（如图 2-1-3），我们发现，深圳的主要特征是开发类岗位需求相较其他城市占比更大（10%），运营岗、产品岗与算法岗需求也较为突出。数据团队的各类岗位都有较大缺口。

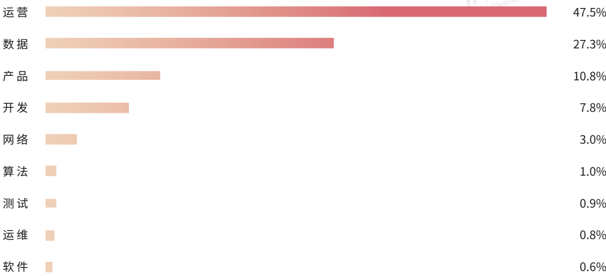
广州则是运营类岗位需求最为旺盛，占比超六成，其他技术性岗位的占比明显较少，数据团队岗位缺口目前偏重运营。

上海的整体结构与深圳类似，但数据类岗位需求更多，达到 33.1%，相应的，运营类岗位需求占比较少，为 42.9%。

北京的数据类岗位需求占比最大，达到北京所有数据相关岗位的 48.4%，产品类也达到 17.2%，算法类岗位需求体量较小；但也是前五名城市中，同类岗位需求占比最大的城市，达到 1.6%，基本呈现技术类与产品类岗位需求为主的职位结构。

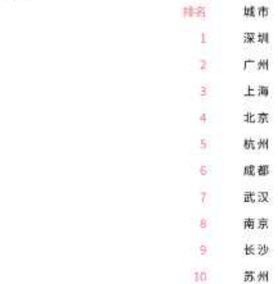
杭州同样与深圳类似，但整体来说，开发、算法等核心技术岗位需求略少。

图2-1-1 数据相关各类在招岗位数量分布



数据来源：各大招聘网站公开发布信息

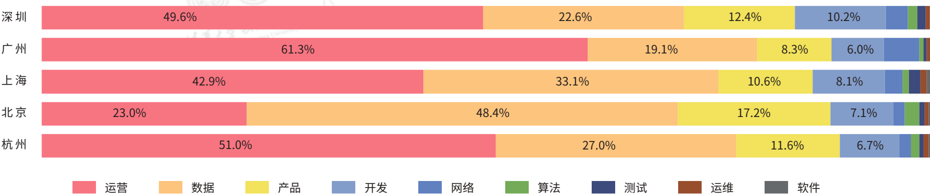
图2-1-2 数据相关在招岗位在各城市分布



排名	城市
1	深圳
2	广州
3	上海
4	北京
5	杭州
6	成都
7	武汉
8	南京
9	长沙
10	苏州

数据来源：各大招聘网站公开发布信息

图2-1-3 典型城市中，数据相关各类在招岗位分布



数据来源：各大招聘网站公开发布信息

2.2 工作经验

3-5 年工作经验人才更吃香

在所有公开招聘职位中，年限要求主要集中在 0 至 5 年，占整体职位数量的 88%。整体来说，数据行业与其他行业类似，不同工作年限的人群组成呈现金字塔结构，而其中，对有一定经验人群的要求比较突出。（如图 2-2-1）

产品、开发、算法岗位更渴望经验丰富的团队领导人

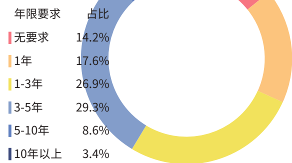
网络、软件、运营、运维、测试五类（如图 2-2-2）岗位的市场需求主要集中在 3 年以下工作经验，占该类岗位需求的六成。而产品、开发、算法岗位则更倾向于更有经验的工作人群，其中产品岗位 3 年以上工作要求的职位占比为 65%，开发为 56%，算法为 49%。在所有数据相关的产品岗位中，要求 5 年以上工作经验的比例同样是所有岗位最高，约为 21%。

由此可以初步看出，产品、开发、算法这些数据行业的核心岗位正在期望可以聘请到能快速带领团队实现业务发展的成熟工作人群。

团队发展周期：广州深圳尚在早期，上海杭州较为均衡，北京偏重经验人群

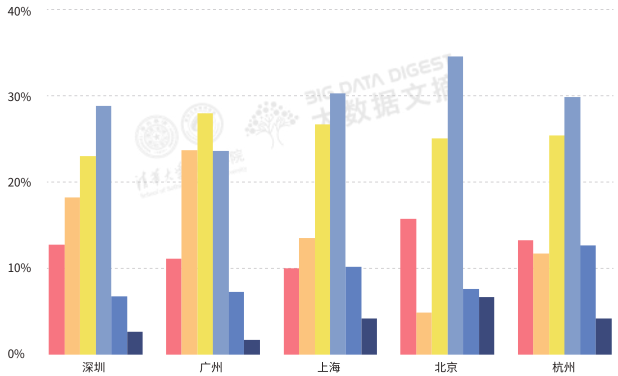
在五个城市中（如图 2-2-3），广州的职位年限要求更偏年轻化，3 年以下工作岗位的数量更多，间接体现其所在地团队的发展成熟度较低。深圳与广州类似，但工作经验要求开始逐渐向 3-5 年阶段过渡，出现逐渐成熟化的市场需求趋势。上海与杭州的招聘职位工作年限结构基本一致，在 1 年、1-3 年、3-5 年、5-10 年四个阶段的分布较为均衡，市场需求更为多元。而北京则出现了针对多年工作经验人群的明显偏好，1 年工作经验占比显著较少，3-5 年与 10 年以上岗位占比较其他城市明显更多。

图2-2-1
数据相关在招岗位
年限分布



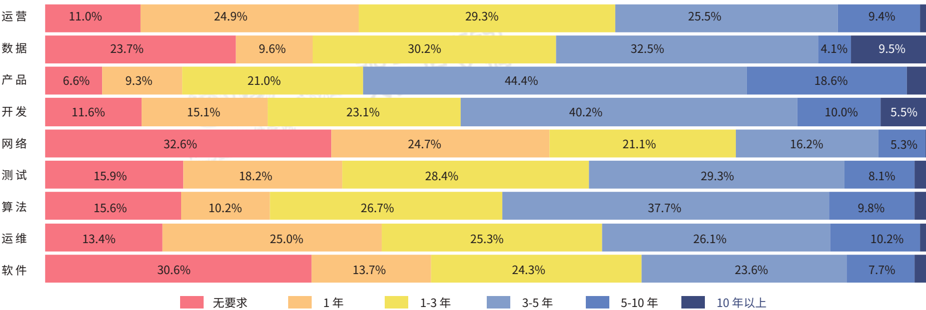
数据来源：各大招聘网站公开发布信息

图2-2-3
数据相关在招岗位
在各典型城市的年限要求



数据来源：各大招聘网站公开发布信息

图2-2-2 数据相关各类在招岗位年限要求



数据来源：各大招聘网站公开发布信息

2.3 学历要求

算法类岗位中，35% 要求硕士及以上学历

数据行业的学历要求中，过半都需要本科以上学历，但对硕士以上学历的要求并不强烈，仅为 2.8%。（如图 2-3-1）

与工作经验要求类似，算法、产品、数据、开发四类核心岗位是对学历要求最高的岗位，本科及以上学历的要求均超过 60%，其中算法岗位的学历要求中本科率最高，本科及以上学历要求达到 97%，硕士及以上学历要求达到 35%。而网络、运营、运维三类职位的学历要求明显偏低，本科及以上学历要求均不超过 50%，其中，网络类职位仅有 18% 求职者获得本科学历。（如图 2-3-2）

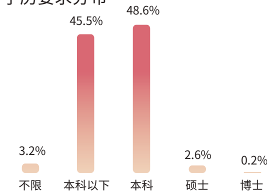


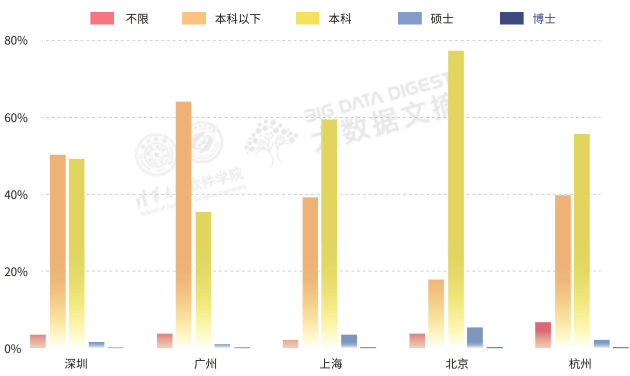
图2-3-1
数据相关在招岗位
学历要求分布

数据来源：各大招聘网站公开发布信息

北京从业者学历要求最高，5% 要求硕士以上学历

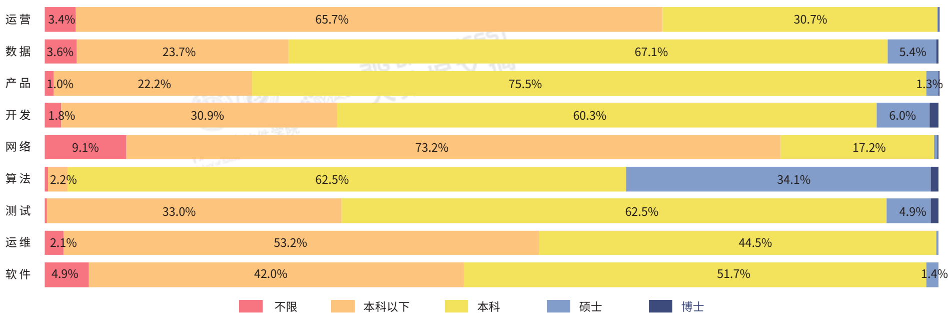
北京对于从业者整体的学历要求最高，本科及以上学历的职位占比达 79%，其中硕士及以上学历要求的占比达 5.2%，同样为五个城市中最高。上海和杭州对于学历的整体要求次之，本科及以上学历占比分别为 60% 和 55%。深圳和广州虽然在招岗位数量最多，但对于学历的整体要求并不高，尤其是广州，本科以下学历要求的比例达到 61%。（如图 2-3-3）

图2-3-3 数据相关在招岗位在各典型城市的学历要求



数据来源：各大招聘网站公开发布信息

图2-3-2 数据相关各类在招岗位学历要求



数据来源：各大招聘网站公开发布信息

2.4 薪资水平

算法博士平均月入4万元，运维、网络、维护类职位薪资较低

数据团队从业者的薪资区间主要集中在5千元至2万元之间，其中5千元至1万元比重更大，占47.7%，1-2万元占比32.6%。5万元以上的高收入群体占比最小，为0.3%。（如图2-4-1）

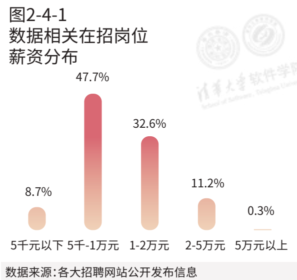
与过去两年趋势相似，算法、产品、开发、数据、研发五类数据行业核心职位仍然占据着高薪岗位前五名，其中，算法岗的平均工资已接近2万。相比之下，运营、运维、网络、维护四类岗位的薪资收入则不甚可观，均在万元以下。（如图2-4-2）

从学历与工作经验角度观察数据团队从业者的工资变化，我们发现，本科学历对于从业者至关重要，决定了其是否能迈过月薪上万的门槛，本科学历比本科以下学历月薪收入高出64%，从9173元跃升至15069元。但本科学历与硕士学历在求职市场上的议价差异并不明显，要求硕士学历的薪资水平仅比本科学历高出2000元。而博士学历则是薪资跃升的最终环节，要求博士学历的岗位对应着接近2.5万元的平均月薪。（如图2-4-3）

在各类职位中，自然是拥有博士学位的算法岗位薪资水平最高，接近4万元。而软件、开发、产品、数据等职位也在硕士或博士要求的职位中，出现了薪资的明显增长。而网络、维护等职位在本科与硕士学历岗位之间却出现了薪资倒挂现象，间接说明该类职位对于学历水平要求并不高。（如图2-4-4）

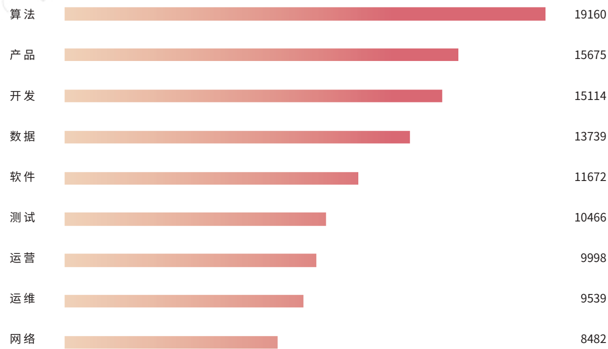
随着工作资历的不断积累，从业者的收入水平也会获得稳步提升，从1年工作经验的7000元水平，在3年左右会突破月薪万元大关，但在5-10年阶段，目前的公开招聘信息中，月薪并没有呈现较快增长，约为2万元。（如图2-4-5）

图2-4-1
数据相关在招岗位薪资分布



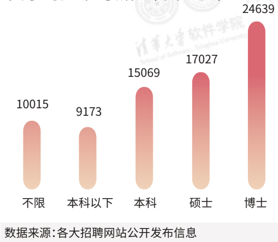
数据来源：各大招聘网站公开发布信息

图2-4-2 数据相关各类在招岗位平均薪资（单位：元）



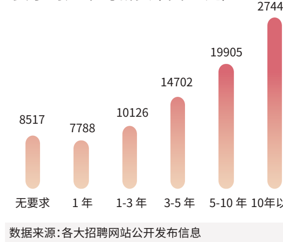
数据来源：各大招聘网站公开发布信息

图2-4-3 数据相关在招岗位各类学历要求对应平均薪资（单位：元）



数据来源：各大招聘网站公开发布信息

图2-4-5 数据相关在招岗位各类年限要求对应平均薪资（单位：元）



数据来源：各大招聘网站公开发布信息

图2-4-4 数据相关各类在招岗位各类学历要求对应平均薪资（单位：元）

	本科以下	本科	硕士	博士
算法	11705	18195	20699	39333
产品	11479	17521	12648	27083
开发	10596	17880	18003	21233
数据	8460	15836	18082	27115
研发	9420	12296	12776	21282
软件	9022	13420	24125	
测试	7463	12284	15549	
运营	9154	12477	11505	22500
运维	8388	11347		
网络	7995	11309	9167	
维护	5938	11432	9000	

数据来源：各大招聘网站公开发布信息

PART 3

数据团队从业者微观洞察

BIG DATA DIGEST 大数据文摘
清华大学软件学院

数据团队从业者并不仅仅是以求职招聘的姿态出现在我们的视野中，他们在实际工作中的思考、问题与困惑，也同样值得关注。因此，我们以数据团队从业者作为对象，进行了一次问卷调查，力求还原以上的一些面向。

3.1工具、技术、方法论

Python 仍是主流数据分析工具

从问卷结果来看，受访者最主要使用的数据分析工具是 Python，Excel 与 MySQL 紧随其后，分别占受访用户群体的 31.8%、22.8% 与 18.6%。（如图3-1-1）

相比往年数据，数据仓库类型的使用工具 Hive、Apache Spark 与 Hadoop MapReduce 的使用频率，有明显的上升，同时 BI 分析工具的使用频率提高了。

Hadoop Hive 是使用最多的大数据平台

在使用的大数据平台以及数据接口上，Hadoop Hive 荣登榜首，占比达到 19.3%，Spark SQL 紧跟其后，占比 13.4%。8.7% 的受访者不确定应该使用何种数据平台和数据接口，所以会考虑不断更换数据平台来进行尝试与对比。此外，使用其他数据平台的受访者行业分布比较分散，均是结合企业本身发展的特点来选择数据平台。（如图3-1-2）

数据整合中最大的挑战：数据的准确性

在数据整合中，29.2% 的用户认为数据的准确性是最大的挑战。数据准确性确实是数据团队在工作中比较难以控制的。其次是数据源分散和数据结构的问题，占比分别为 26.8% 和 22.5%。此外，数据集成、数据量和数据速率对于众多数据团队来说也是不小的挑战，均超过 10%。（如图3-1-3）

业务分析与数据分析成为通用必会的技能

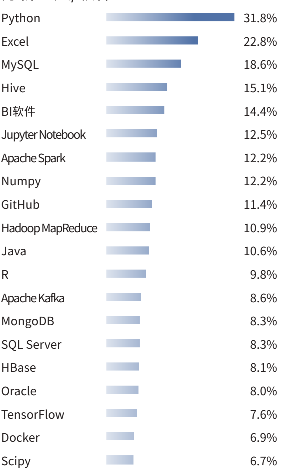
根据领英数据显示，在全球范围内，数据人才排名靠前的两项通用技能是业务分析和数据分析，分别是 57% 和 41%，数据分析与业务数据一直都是紧密联系的。相比之下，解析学、SQL 和 Python 技能的使用频率不容忽视，熟练掌握解析学对解读数据之间的联系会有帮助，掌握SQL 可以高效地查询与汇总数据，而 Python 则是因为有丰富多样的数据库，可以顺利完成数据处理、数据分析与可视化工作。（如图3-1-4）

在中国范围内，数据人才的通用技能前二十并无二致。略有不同的是，国内数据人才的通用技能更强调的是编程语言软件的熟练使用程度，比如 SQL、Python 和 R。与全球市场的结果相比较，国内市场使用像解析学、数据挖掘以及机器学习的数据分析基础技能比较多，而且三者的占比几乎是不相上下。国内市场更多的是遵循业务数据的分析、数据分析工具的掌握和数据解读的基础技能这样的通用技能排序。（如图3-1-5）

数据可视化技能正在成为全球增长最快的技能

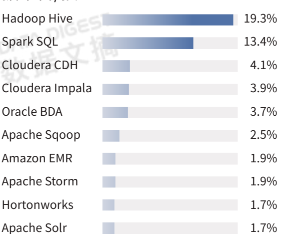
领英数据显示，数据分析技能和数据可视化技能是过去一年在全球数据人才中增长最快的技能，前者增长速度约为 150%，后者增速同样超过 100%，Microsoft Power BI 在过去一年的增长速度为 94.4%，是全球市场的新兴技能排序中首位增长速度最快的分析工具。（如图3-1-6）

图3-1-1 数据团队从业者常用的数据分析工具/软件 TOP20



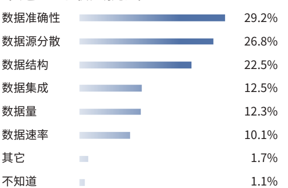
数据来源：大数据文摘读者问卷调研

图3-1-2 数据团队从业者常用的大数据平台/接口



数据来源：大数据文摘读者问卷调研

图3-1-3 数据团队从业者在数据整合中遇到的最大挑战



数据来源：大数据文摘读者问卷调研

人工智能、数据科学与数据管理技能同样呈现较快增长速度。在过去一年中，人工智能与数据科学对数据分析的影响比较大，数据量的日益增长也促使了数据管理技能的高速发展。（如图3-1-7）

在国内市场，过去一年中增长最快的技能为数据科学，增长速度为175.4%，相较全球范围，国内对于数据科学技能的需求更旺盛。其他技能的需求程度与国际整体较为相似。

图3-1-4 领英全球范围内数据人才通用技能TOP20

排名	技能	占比
1	业务分析	57.1%
2	数据分析	41.1%
3	SQL	28.3%
4	解析学	24.3%
5	Python	17.5%
6	分析能力	15.4%
7	商业智能	14.3%
8	R	13.1%
9	数据科学	11.8%
10	机器学习	11.7%
11	需求分析	11.2%
12	数据库	11.1%
13	金融	10.7%
14	数据挖掘	10.3%
15	Tableau	9.9%
16	Microsoft SQL Server	9.5%
17	Java	9.4%
18	软件开发	9.4%
19	业务流程优化	8.6%
20	统计	8.3%

数据来源：领英

图3-1-5 领英中国范围内数据人才通用技能TOP20

排名	技能	占比
1	数据分析	58.3%
2	业务分析	28.8%
3	SQL	24.1%
4	Python	21.5%
5	R	14.4%
6	解析学	13.1%
7	数据挖掘	12.6%
8	机器学习	10.5%
9	Tableau	8.6%
10	金融	8.2%
11	IBM SPSS	7.3%
12	Java	6.5%
13	MATLAB	6.3%
14	商业智能	6.2%
15	SAS	5.7%
16	金融分析	5.3%
17	数据库	5.3%
18	数据科学	5.2%
19	统计	5.1%
20	MySQL	5.0%

数据来源：领英

图3-1-6 领英全球数据人才市场增长最快技能TOP20

排名	技能	占比
1	数据分析	149.2%
2	数据可视化	111.5%
3	Microsoft Power BI	94.4%
4	人工智能	81.7%
5	数据科学	73.1%
6	数据管理	54.2%
7	Tableau	46.5%
8	Jira	44.8%
9	Python	44.1%
10	大数据	43.8%
11	机器学习	41.7%
12	业务分析	39.6%
13	统计数据分析	38.4%
14	数据建模	33.0%
15	统计建模	32.7%
16	信息技术	30.1%
17	解析学	29.2%
18	数据挖掘	27.7%
19	Scrum	27.0%
20	Hadoop	26.0%

数据来源：领英

图3-1-7 领英中国数据人才市场增长最快技能TOP20

排名	技能	占比
1	数据科学	175.4%
2	数据分析	165.3%
3	业务需求	135.9%
4	数据可视化	131.4%
5	Microsoft Power BI	101.3%
6	商业分析	79.6%
7	大数据分析	71.4%
8	深度学习	70.7%
9	数据管理	68.5%
10	人工智能	64.9%
11	信息技术	61.3%
12	Tableau	57.8%
13	统计数据分析	42.9%
14	机器学习	41.5%
15	技术支持	38.5%
16	Python	36.3%
17	数据建模	35.9%
18	Visio	35.8%
19	大数据	35.3%
20	Apache Spark	34.4%

数据来源：领英

BIG DATA DIGEST 大数据文摘
清华大学软件学院

3.2 自我认知

从业者转行意愿低，多数职业规划方向为专家或决策者

超过半数的受访者对目前的职业状态感到满意，表达不满的受访者比例约为 15%（如图 3-2-1）。对于职业发展的正向期待，从他们的五年规划中也可见一斑。

对于受访者来说，在他们的五年规划里，数据人才更愿意深挖数据方向的工作，意愿达到 38.8%，其次有 15.1% 希望升职成为企业决策层。还有一部分数据人才更希望转型做产品经理或者尝试创业，占比分别为 12% 和 11.5%。希望维持目前工作不变的比例约为 11.5%。真正考虑转行的从业者数量仅为 1.4%。整体而言，受访者对于自己的职业规划都集中在决策者和专家两条道路上。（如图 3-2-2）

3.3 工作生活

各类书籍与官方文档是从业者主要的学习渠道

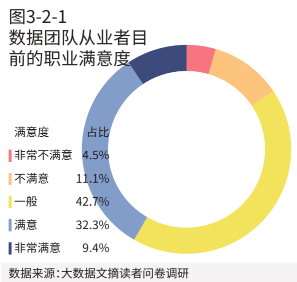
受访者中，主流的业余学习习惯都与书本和官方文档有关，偏向于官方的体系化知识获取，占比均超过 6 成。遇到具体问题，也会有 36.2% 的用户使用官方在线帮助服务或询问朋友。但国内从业者对于如 StackOverflow 类的在线问答社区并不热衷，仅有 26.8% 的受访者会采用这一形式。（如图 3-3-1）

两成受访者为开源社区做出过贡献，相应的，约有近八成受访者没有为开源社区做出贡献。（如图 3-3-2）

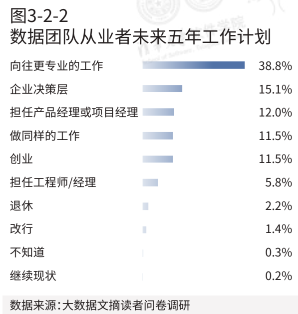
3.4 数据应用场景

数据团队工作成果主要作用于运营、市场等业务方向

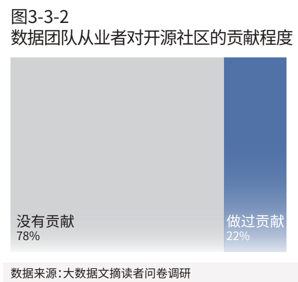
从问卷调研结果来看，目前企业中的数据团队工作成果主要作用于运营、销售、市场、客户服务等直接业务层面与战略决策方向，占比均超过 30%。而直接对接外部客户形成数据团队产品落地的情况相对较少，仅为 12.6%。（如图 3-4-1）



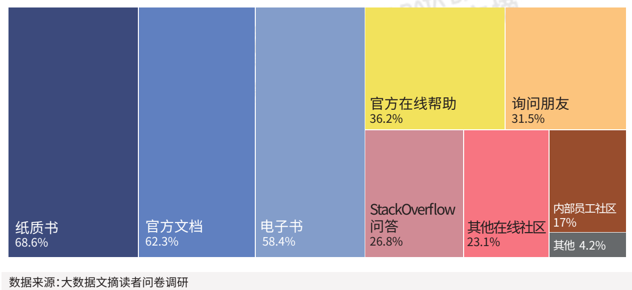
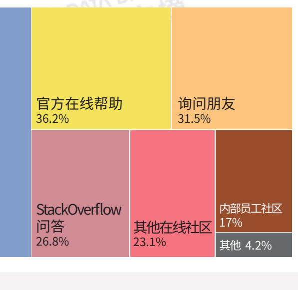
数据来源：大数据文摘读者问卷调研



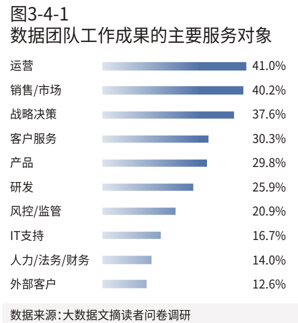
数据来源：大数据文摘读者问卷调研



数据来源：大数据文摘读者问卷调研

数据来源：大数据文摘读者问卷调研



数据来源：大数据文摘读者问卷调研

3.5 团队前景

超过半数受访者对未来一年公司对数据团队的投入情况表示乐观

在团队未来一年的资金健康状况上，54% 的受访者认为所在机构在未来一年会对数据团队增加投资，推动业务进一步发展。19% 的受访者则认为企业对团队的投入不会增加或可能出现减少。（如图 3-5-1）

3.6 大数据与人工智能问题

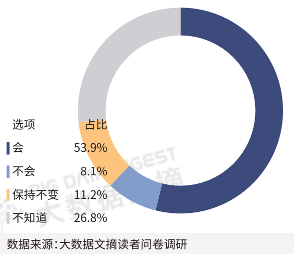
八成受访者对人工智能技术感到兴奋

对于大数据与人工智能这一经久不衰的讨论方向，受访者中超过八成对于人工智能技术的到来表示欢迎和兴奋，他们更期待人工智能可以增加工作中的自动化程度，通过算法辅助决策并提升公平性。（如图 3-6-1 及 3-6-2）

12.6% 的受访者则表示了更多担忧，担忧的方向集中在人工智能是否会超越人类、算法将如何影响决策等方面。由此说明，人工智能在日常工作中的利弊优劣还在逐渐被从业者熟悉的阶段，而至于技术未来将如何左右行业发展，也将由这些人的行动决定。（如图 3-6-3）

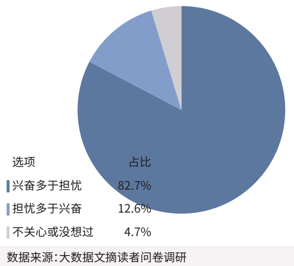
而面对使用大数据与人工智能技术可能带来的道德问题，有 34.5% 的受访者认为政府 / 监管机构应该负主要责任，同时也有 30.2% 的人认为所在公司的经理以及管理者应该负主要责任，17.6% 认为技术人员 / 研究者应该负主要责任。（如图 3-6-4）

图3-5-1数据团队从业者所在机构对数据团队的未来投资



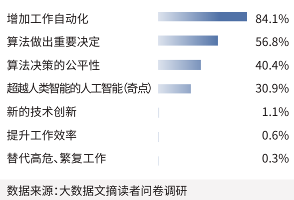
数据来源：大数据文摘读者问卷调研

图3-6-1
数据团队从业者如何看待人工智能



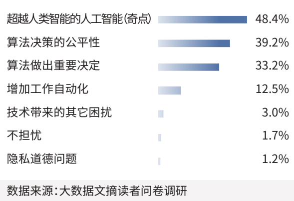
数据来源：大数据文摘读者问卷调研

图3-6-2
数据团队从业者对人工智能的期待



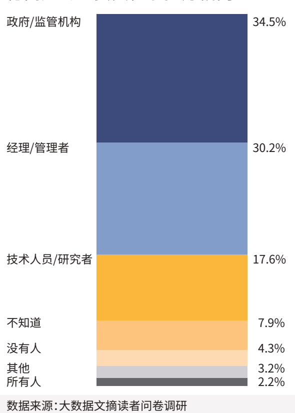
数据来源：大数据文摘读者问卷调研

图3-6-3
数据团队从业者对人工智能的担忧



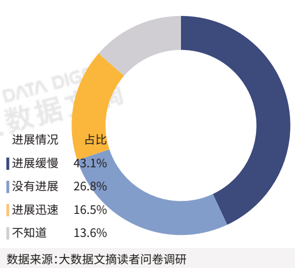
数据来源：大数据文摘读者问卷调研

图3-6-4 数据团队从业者对于技术道德问题的主要责任承担方偏向



数据来源：大数据文摘读者问卷调研

图3-6-5 数据团队从业者所在团队过去一年人工智能技术进展情况



数据来源：大数据文摘读者问卷调研

七成数据团队中的 AI 技术进展缓慢

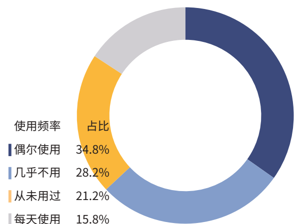
调查结果显示，与对于人工智能的乐观情绪相比，数据团队中的人工智能技术发展并不容乐观。在过去的一年里，43.1% 的受访者认为团队中的 AI 技术发展缓慢，甚至有 26.8% 的受访者认为没有进展。只有 16.5% 的受访者认为进展迅速。（如图 3-6-5）

相应的，受访者表示，在日常工作中使用 AI 相关技术（深度学习、自

然语言处理、计算机视觉等技术）的频率并不高，仅有 15.8% 表示几乎每天都在使用相关技术，34.8% 仅偶尔使用，近一半受访者几乎不使用或者从未使用过相关技术。（如图 3-6-6）但仍有高达 76% 的受访者表示，会考虑在未来的工作场景中频繁使用相关技术。（如图3-6-7）

在具体应用场景上，15.3% 的受访者认为应该把 AI 技术应用到精准推荐，9.8% 的受访用户则认为应该应用到云计算，此后则是聊天机器人、人脸识别、智能硬件、图像识别、虚拟现实等场景。（如图3-6-8）

图3-6-6 数据团队从业者工作中用到人工智能技术的频率



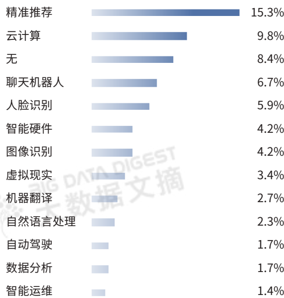
数据来源：大数据文摘读者问卷调研

图3-6-7 数据团队从业者在未来一年是否考虑在工作中更频繁使用人工智能技术



数据来源：大数据文摘读者问卷调研

图3-6-8 数据团队从业者所在团队过去一年人工智能技术进展



数据来源：大数据文摘读者问卷调研

PART 4

全球数据团队观察

BIG DATA DIGEST 大数据文摘
清华大学软件学院

对于数据团队或数据产业的观察，绝不应仅局限于在中国范围内，观察全球前沿国家发展现状同样至关重要。

领英基于全球会员的全样本数据，从数据库中筛选出符合要求的数据职位名称，并根据这些职位名称筛选出符合要求的数据从业者。根据领英的全球职位数据，我们把视野从中国本土扩展到全球，着重选取了美国、印度、英国三个数据人才分布最集中的国家，将这些国家的数据人才的相关数据与中国对比，观察数据团队及数据从业者现状。

4.1 学历分布和毕业院校

中国数据团队从业者硕士学历比例超全球均值，英国博士学历从业者最多

随着数据行业近年来的发展，从全球范围看，招聘市场对数据人才表现出极高的需求，全球数据从业者也在稳步增长，过去一年增长率 4.4%。从全球数据从业者的学历分布来看，数据人才普遍受教育程度高，98.5% 拥有本科或以上学历，其中持本科学历的数据从业者最多，占比 44%，持硕士学位者次之，占 36.4%，持 MBA 学位者占 12.1%，持博士学位者占比 6%。（如图 4-1-1）

具体观察各个国家，领英站内印度本科及以上学历水平的从业人员占比最高，几乎所有数据行业从业人员都拥有本科或以上学历，英国拥有最多博士学历的数据从业者，占比达 8.3%，美国博士学历的数据从业者占比 7.4%。

与以上三国对比，中国的硕士学历从业者占比最多，达 51.2%，本科次之，占比 41%，但 MBA 与博士学历从业者较少，分别为 3.6% 和 3.1%。

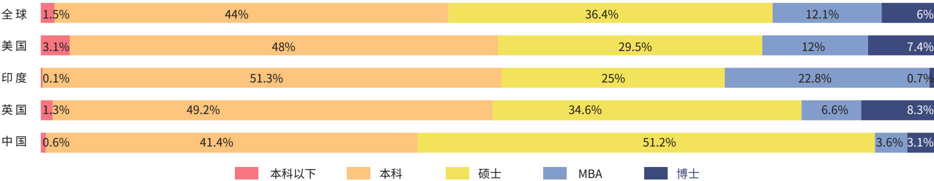
全球范围内，培育数据团队从业者最多的院校以美国和印度高校为主。按照从事数据职业的毕业生人数排序，前十大院校中印度占 5 位，美国占 4 位，加拿大占 1 位，印度孟买大学拥有全球最多从事数据行业的毕业生，前三中美国凤凰城大学和美国加州大学伯克利分校分别位列第二和第三，加拿大多伦多大学排名第七。

纵观 2016 年到 2019 年间，新毕业的数据团队从业者的学历水平，硕士学历占比最多，达 43.2%，本科学历降至 38.7%，MBA 学历占比 11.5%，博士学历占比为 5.8%，本科以下学历占比为 0.8%。

印度高校数据人才实力崛起，美国高校整体优势未被打破

过去四年间，培育新晋数据人才最多的前十大院校排名也有着显著变化，印度尼赫鲁科技大学从 2017 年起连续三年成为培育数据从业者最多的学校。印度中央政府学校保持逐年上升趋势，从十名以外稳步上升至第二名。加州大学伯克利分校、加拿大多伦多大学、新加坡国立大学三所大学紧随其后。除印度和美国高校外，巴西番达曹格图里奥瓦尔加斯大学连续四年上榜，荷兰阿姆斯特丹大学从 2016 年的第八名跌出前十名。（如图 4-1-2）

图4-1-1 领英站内各典型国家数据从业人员学历分布



数据来源：领英

图4-1-2 领英站内数据团队从业者毕业院校 TOP10（2016-2019）

排名	2016 年	国家	2017 年	国家	2018 年	国家	2019 年	国家
1	美国哥伦比亚大学	美国	印度尼赫鲁科技大学	印度	印度尼赫鲁科技大学	印度	印度尼赫鲁科技大学	印度
2	美国加州大学伯克利分校	美国	美国加州大学伯克利分校	美国	美国加州大学伯克利分校	美国	印度中央政府学校	印度
3	加拿大多伦多大学	加拿大	新加坡国立大学	新加坡	加拿大多伦多大学	加拿大	美国加州大学伯克利分校	美国
4	美国佐治亚理工学院	美国	加拿大多伦多大学	加拿大	印度中央政府学校	印度	加拿大多伦多大学	加拿大
5	新加坡国立大学	新加坡	美国加州大学圣地亚哥分校	美国	巴西番达曹格图里奥瓦尔加斯大学	巴西	美国弗吉尼亚大学	美国
6	美国斯坦福大学	美国	美国德克萨斯大学奥斯汀分校	美国	美国加州大学圣地亚哥分校	美国	美国德克萨斯A&M大学	美国
7	美国华盛顿大学	美国	巴西番达曹格图里奥瓦尔加斯大学	巴西	美国德克萨斯大学奥斯汀分校	美国	美国罗格斯大学	美国
8	荷兰阿姆斯特丹大学	荷兰	美国华盛顿大学	美国	新加坡国立大学	新加坡	新加坡国立大学	新加坡
9	巴西番达曹格图里奥瓦尔加斯大学	巴西	印度中央政府学校	印度	美国弗吉尼亚大学	美国	美国西北大学	美国
10	美国西北大学	美国	美国宾夕法尼亚州立大学	美国	美国华盛顿大学	美国	巴西番达曹格图里奥瓦尔加斯大学	巴西

数据来源：领英

4.2 主流专业分布

全球计算机与数学相关专业为主流，英国出现部分心理学化学专业从业者

根据领英数据，统计全球范围内不同专业背景的数据人才数量，可以得出培养数据从业者的十大大主流专业：计算机与数学相关专业在前十名中占据五席，计算机科学排名第一，信息技术第四，数学第五，统计学第七，计算机科学与技术第九。其他上榜专业中，除排名第二的工商管理与排名第八的市场营销外，都是要求较强数理背景的专业，如经济学、金融、物理学等。（如图 4-2-1）

对比各国，美国主流专业排名第一的是工商管理，计算机科学、信息技术分别排名第二和第六，会计学进入美国主流专业前十。印度的主流专业更多偏向工科，除了排名第一的计算机软件和排名第七的计算机工程外，电气、电子和通信工程排名第三，电气电子工程排名第十，与数据行业不直接相关的商务贸易排名第四。在英国，数学专业排名第一，计算机软件排名前三，计算机和信息科学及支持服务排名第七，此外，心理学和化学两个相关性更弱的专业也进入榜单，分列第九和第十，从侧面反映了英国数据团队从业者的多元化。与其他三个国家相比，中国主流专业排名第一的是统计学，计算机方向的计算机科学和计算机软件工程分别排名第二和第八，金融学排名第三，数学、管理信息系统、市场营销等专业亦上榜。（如图 4-2-2）

4.3 数据人才集中分布的行业

互联网与管理咨询行业数据人才需求与增长率双高

观察数据人才最集中的十大行业，我们可以看到，信息技术与服务、金融服务和计算机软件占据前三名，它们对于数据人才的需求程度也处在较高水平。此外，互联网和管理咨询同时有着极高的人才需求和人才供应，属于数据相关行业中造血功能更强后劲更足的领域，其年增长率分别为 10.7%、9.8%，银行和电信行业的增长率与招聘需求则较弱，行业横向对比的发展潜力处于劣势。（如图 4-3-1）

在美国，信息技术与服务、金融服务都是数据人才最集中的两大行业，管理咨询、互联网和高等教育的年增长率分别为 12.3%、9.5% 和 7.9%，其中管理咨询和互联网的招聘需求极高，高等教育则为中等，保险虽是数据人才集中行业第四，但其招聘需求比较低，市场需求趋向饱和。

与美国类似，在印度，信息技术与服务和金融服务为数据人才最集中的行业，其中金融服务对未来数据人才有着极高的需求，而信息技术与服务为高。互联网和管理咨询的数据人才年增长率达到了 14% 和 12%，有着极高的招聘需求，而外包 / 离岸外包和零售的招聘需求仅为中等。

英国的金融服务尽管是数据人才最集中的行业，但其年增长率却为 -2.1%，招聘需求也为中等，该市场已经饱和，同时银行、保险和零售的年增长率同为负值，但与招聘需求低的银行不同，保险和零售对数据人才的招聘需求表现为高和极高。除此之外，政府事务、高等教育和公共事业对数据人才的需求低。

在中国，互联网、信息技术与服务和金融领域为数据人才最集中的行业，相对应的年增长率分别为 5.7%、2% 和 2.5%，对数据人才的招聘需求也都为高或较高。零售的年增长率出现负值，为 -0.8%，但对数据人才的招聘需求高，说明行业正在力图引进更多数据人才改变目前状况。（如图 4-3-2）

图4-2-1 领英站内全球范围内数据从业者主流专业

排名	专业名称
1	计算机科学
2	工商管理
3	经济学
4	信息技术
5	数学
6	金融
7	统计学
8	市场营销
9	计算机科学与技术
10	物理学

数据来源：领英

图4-2-2 领英站内各典型国家数据从业者主流专业

排名	美国	印度	英国	中国
1	经济学	计算机科学	数学	统计学
2	工商管理	电气、电子和通信工程	经济学	金融
3	计算机科学	信息技术	工商管理	经济学
4	统计学	机械工程	计算机科学	工商管理
5	金融	计算机工程	物理	数学
6	信息技术	金融	统计学	计算机科学
7	数学	电子电气工程	心理学	会计学
8	管理信息系统	市场营销	金融	市场营销
9	市场营销	商务贸易	信息技术	管理信息系统
10	心理学	工商管理	分析与功能分析	计算机软件工程

数据来源：领英

4.4 数据人才的跨行业流动

各行业数据人才涌入互联网，高等教育行业数据人才流失严重

与其他专业人才相似，数据人才同样会在不同行业之间流动。数据人才最集中的十大行业中，不同行业对数据人才的吸引力指数也有所区别。

（注：行业人才吸引力指数 = 流入指定行业的人才 / 流出该指定行业的人才）

从总的趋势上看，从高等教育进入到各行业的数据人才占比排名都是第一，在与计算机软件、医院与护理和互联网的互动中，数据人才呈现出单向流出的趋势，即从高等教育流出到这些行业中，而在与信息技术与服务和金融服务的互动中，数据人才则为双向流动。

图4-3-1 领英站内全球范围内数据人才最集中的十大行业增长率及招聘需求程度

数据人才数量排名	行业名称	增长率	招聘需求
1	信息技术与服务	5.1%	高
2	金融服务	3.8%	极高
3	计算机软件	6.6%	极高
4	银行	4.4%	中等
5	医院与护理	6.6%	极高
6	保险	3.8%	高
7	互联网	10.7%	极高
8	管理咨询	9.8%	极高
9	电信	1.7%	中等
10	高等教育	6.2%	高

数据来源：领英

互联网行业保持了较高的吸引力，在数据人才的流动上，信息技术与服务、金融服务、管理咨询、通信和高等教育行业中的数据人才多流入到互联网。

在计算机软件、银行、医院与护理、保险行业和互联网行业，数据人才都是从其他行业流入，而高等教育的数据人才流出现象严重，其中互联网的吸引力指数最高，达到 6.27，金融服务次之，为 5.20。（如图 4-4-1）

图4-4-1 领英站内全球范围内数据人才在各行业流入流出情况



注：行业人才吸引力指数 = 流入指定行业的人才/流出该指定行业的人才

数据来源：领英

图4-3-2 领英站内各典型国家数据人才最集中的十大行业增长率及招聘需求程度

数据人才数量排名	美国	增长率	招聘需求	印度	增长率	招聘需求	英国	增长率	招聘需求	中国	增长率	招聘需求
1	信息技术与服务	3.6%	极高	信息技术与服务	4.4%	高	金融服务	-2.1%	中等	互联网	5.7%	极高
2	金融服务	3.3%	极高	金融服务	5.9%	极高	信息技术与服务	2.6%	高	信息技术与服务	2.0%	极高
3	医院与护理	3.2%	中等	计算机软件	7.3%	极高	银行	-5.6%	低	金融	2.5%	高
4	保险	2.4%	低	互联网	14.0%	极高	政府事务	6.0%	低	计算机软件	1.2%	高
5	计算机软件	5.3%	极高	管理咨询	12.0%	极高	计算机软件	4.1%	极高	管理咨询	4.2%	极高
6	高等教育	7.9%	中等	银行	5.0%	高	高等教育	2.8%	低	通信	0.3%	中等
7	互联网	9.5%	极高	市场营销与广告	8.3%	高	保险	-2.7%	高	银行	3.9%	中等
8	管理咨询	12.3%	极高	通信	1.9%	高	管理咨询	3.5%	极高	制药	8.6%	极高
9	零售	3.9%	极高	外包/离岸外包	5.2%	中等	公共事业	-0.8%	低	汽车	1.9%	高
10	银行	4.0%	中等	零售	4.2%	中等	零售	-1.1%	极高	零售	-0.8%	高

数据来源：领英

4.5 数据人才的主流职位和增长最迅速的数据相关职位

全球范围商业分析师最为主流，中国数据分析师更吃香

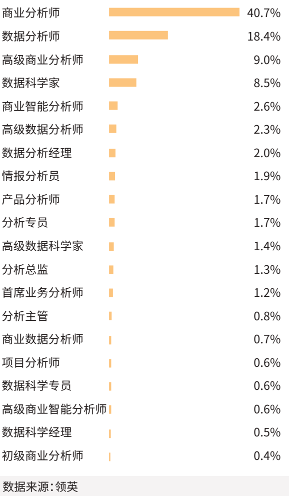
在全球范围内，数据人才在选择职业时，商业分析师为最受欢迎的职位，占比达到 40.7%，比排名第二的数据分析师的占比 18.4% 高出了一倍多，高级商业分析师排名第三，占比 9%。综合来说，商业分析师职业路径是数据人才最主流的成长路径，这一特征普遍存在于美国、印度和英国的数据人才就业选择中。（如图 4-5-1）

在这份职位选择榜单中，数据人才对于商业分析师和数据分析师类工作的青睐可见一斑，几乎所有职位都与分析师相关，此外则是对于数据科学背景要求更高的数据科学家职位。

在美国，商业分析师占比达到 33.6%，数据分析师占比 17.6%，高级商业分析师占比 11.2%，这也是美国数据人才职位头衔占比唯三超过 10% 的职位。印度数据人才对于商业分析师职位的青睐最为明显，占比超过一半，达到 54.7%，比排名第二的数据分析师所持占比 17.3% 高出两倍。在英国，商业分析师超过四成。

相较于商业分析师，中国数据人才则更偏向于数据分析师类职位，后者在数据人才的职位头衔占比上最多，达到 46.9%，均高于其他三个国家同一职位头衔的一倍，排名第二的商业分析师占比 23.3%。（如图 4-5-2）

图4-5-1 领英站内全球范围内数据人才主流岗位



数据来源：领英

数据科学相关职业增长最快

在数据团队从业者的职位变化上，根据领英平台数据，在过去一年，按照职位的增长率计算，在全球范围内增长最迅速的数据相关职位是数据科学经理，增速达到 18.1%，高级数据科学家增速排名第二，为 17.2%，数据科学家排名第三，为 16.1%。排名前九位的相关职位增速均超过 10%。（如图 4-5-3）

过去一年中，在美国，高级商业分析师增速最快，高达 48.7%，比排名第二的咨询商业分析师增速快了两倍左右，除高级商业分析师外，排

图4-5-2 领英站内各典型国家数据人才主流岗位

美国	增速	印度	增速	英国	增速	中国	增速
商业分析师	33.6%	商业分析师	54.7%	商业分析师	41.3%	数据分析师	46.9%
数据分析师	17.6%	数据分析师	17.3%	数据分析师	17.2%	商业分析师	23.3%
高级商业分析师	11.2%	高级商业分析师	7.7%	高级商业分析师	9.6%	分析专员	6.8%
数据科学家	7.9%	数据科学家	6.9%	数据科学家	6.8%	高级数据分析师	4.2%
智能分析师	4.0%	高级数据分析师	2.2%	首席商业分析师	3.3%	数据科学家	3.5%
高级数据分析师	3.2%	首席商业分析师	1.5%	商业智能分析师	2.5%	高级数据分析师	3.4%
分析总监	2.7%	分析经理	1.2%	分析经理	2.2%	分析总监	2.8%
分析经理	2.6%	分析专员	1.1%	高级数据分析师	2.0%	分析经理	1.8%
商业智能分析	2.2%	助理商业分析师	1.0%	智能分析师	1.8%	商业数据分析师	1.8%
高级数据科学家	2.0%	产品分析师	0.9%	分析主管	1.4%	产品分析师	0.9%
分析专员	1.9%	高级数据科学家	0.8%	高级数据科学家	1.3%	商业智能分析师	0.6%
产品分析师	1.6%	商业智能分析师	0.7%	分析专员	1.3%	数据挖掘分析师	0.6%
首席业务分析师	1.1%	数据科学实习生	0.5%	产品分析师	1.2%	高级数据科学家	0.5%
商业数据分析师	0.8%	数据科学专员	0.5%	项目商业分析师	1.0%	首席商业分析师	0.3%
高级商业智能分析师	0.7%	数据科学经理	0.4%	分析总监	0.9%	商业分析实习生	0.3%
数据科学经理	0.7%	首席数据科学家	0.3%	外包商业分析师	0.7%	智能分析师	0.3%
高级分析经理	0.6%	分析主管	0.3%	商业数据分析师	0.7%	数据科学专员	0.2%
数据科学总监	0.6%	高级产品分析师	0.3%	初级商业分析师	0.6%	项目商业分析师	0.2%
高级产品分析师	0.6%	商业数据分析师	0.3%	高级商业智能分析师	0.6%	数据科学经历	0.2%
数据科学专员	0.5%	高级分析经理	0.2%	首席数据科学家	0.5%	分析主管	0.2%

数据来源：领英

名前十的相关职位增速均达到 10% 以上，且差距较小。

在印度，数据科学经理增速最快，达到了 31.3%，与排名第二和第三的高级数据科学家和智能分析师所持占比 24.9% 和 24.4% 相差较小，印度整体增速都较为平均，相邻职位之间的增速相差较小，增速排名前 12 的职位均达到 10% 以上，排名前三的职位增速达到 20% 以上。

英国的整体趋势与印度类似，排名前后的职位增速差距不大，排名前十的职位增速均达到 10% 以上，其中数据科学主管排名第一，增速达到 25.4%，初级商业分析师、高级数据科学家、数据科学专员、数据科学经理的职位增速都超过了 20%。

中国的整体增速趋势仍然较为平均，增速最快的是数据科学经理，达到 29.2%，排名第二的业务控制分析师增速 22.2%，增速排名前七的职位均达到 10% 以上。（如图 4-5-4）

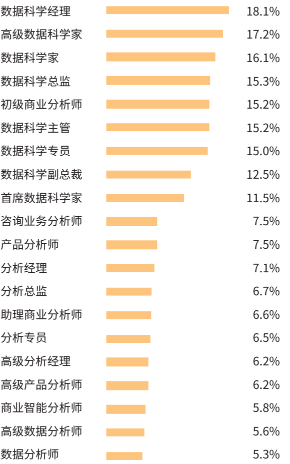
4.6 数据从业者集中分布的地区

在数据从业者集中分布的城市和地区上看，全球范围内的数据人才多集中分布在较发达城市和地区，这也与数据人才集中毕业院校的城市和地区相吻合。（如图 4-6-1）

在全球数据从业者最集中的十大地区中，美国五个地区上榜，大纽约大地区排名第一，旧金山湾区、华盛顿特区 - 巴尔的摩地区、大波士顿地区和大洛杉矶地区分别排名第四、第六、第九和第十。美国对于数据人才的吸引力也较强，除大洛杉矶地区外，其余地区的年增长率均超过 4%，华盛顿特区 - 巴尔的摩地区和大波士顿地区的年增长率达到 5.2%。对于数据人才体现出高需求。

近年数据人才培养能力崛起的印度则有三个地区上榜，大班加罗尔地区排名第二，大德里地区和孟买都市区分别排名第五和第七，且三个地区的年增长率相差不大，均为 5%

图4-5-3 领英站内全球范围内增长最快的数据相关岗位



数据来源：领英

图4-5-4 领英站内各典型国家增长最快的数据相关岗位

美国	增速	印度	增速	英国	增速	中国	增速
高级商业分析师	48.7%	数据科学经理	31.3%	数据科学主管	25.4%	数据科学经理	29.2%
咨询商业分析师	14.7%	高级数据科学家	24.9%	初级商业分析师	23.9%	业务控制分析师	22.2%
数据科学家	14.4%	智能分析师	24.4%	高级数据科学家	22.5%	数据科学主管	18.8%
数据科学总监	14.2%	数据科学家	17.7%	数据科学专员	21.4%	初级商业分析师	16.7%
高级数据科学家	14.1%	数据科学专员	16.8%	数据科学经理	20.2%	高级商业智能分析师	15.6%
数据科学主管	13.9%	产品分析师	15.8%	助理商业分析师	18.5%	数据科学总监	13.8%
数据科学经理	13.8%	数据科学总监	15.8%	数据科学总监	16.3%	区域商业分析师	12.5%
初级商业分析师	12.3%	咨询商业分析师	15.6%	数据科学家	16.3%	助理商业分析师	9.1%
数据科学副总裁	11.2%	数据科学副总裁	15.5%	数据科学主管	15.3%	商业分析实习生	8.8%
数据科学专员	10.0%	高级产品分析师	14.8%	产品分析师	11.3%	智能分析师	8.5%
首席数据科学家	8.7%	首席数据科学家	13.5%	高级产品分析师	9.8%	高级数据科学家	7.2%
产品分析师	8.3%	分析总监	12.8%	高级商业智能分析师	9.6%	数据科学家	7.0%
分析总监	7.0%	商业智能分析师	9.7%	分析专员	8.5%	首席商业分析师	6.7%
分析主管	7.0%	初级商业分析师	9.4%	首席数据分析师	8.1%	数据科学专员	4.3%
分析经理	6.6%	首席数据分析师	8.7%	高级分析经理	7.2%	数据挖掘分析师	4.2%
高级数据分析师	5.7%	分析专员	7.8%	分析经理	6.6%	商业数据分析师	4.1%
高级分析经理	5.0%	分析经理	7.3%	数据分析师	6.4%	商业智能分析师	4.1%
助理商业分析师	4.7%	分析主管	7.3%	高级数据分析师	6.2%	首席数据分析师	4.1%
高级产品分析师	4.3%	助理商业分析师	7.3%	分析主管	5.2%	分析主管	3.8%
分析专员	4.1%	数据科学主管	7.1%	分析总监	4.8%	首席数据科学家	3.7%

数据来源：领英

左右，但印度整体市场对数据人才的招聘需求不高。

此外的集中分布区域还有英国的英格兰大都会地区（排名第三）和巴西的圣保罗地区（排名第八），英国的年增长率为 1.7%，为榜单最低，巴西年增长率 7.8%，为榜单最高。同时，英国体现出极高的数据人才招聘需求，而巴西的数据人才需求仅为中等。

图4-6-1 领英站内数据从业者全球分布情况



排名	地区	国家	增长率	招聘需求
1	大纽约大地区	美国	4.1%	极高
2	大班加罗尔地区	印度	5.0%	中等
3	英格兰大都会地区	英国	1.7%	极高
4	旧金山湾区	美国	4.6%	极高
5	大德里地区	印度	4.9%	低
6	华盛顿特区-巴尔的摩地区	美国	5.2%	高
7	孟买都市区	印度	4.9%	低
8	圣保罗地区	巴西	7.8%	中等
9	大波士顿地区	美国	5.2%	极高
10	大洛杉矶地区	美国	3.4%	极高

数据来源：领英

PART 5

疫情中的数据行业

清华大学软件学院 大数据文摘 BIG DATA DIGEST

毫无疑问，2020 年突如其来的疫情，对所有企业都带来了影响。根据艾瑞咨询的统计数据，在疫情最严重的二月份，29.6% 的中小企业营收下降 50% 以上，85% 的中小企业靠现金只能维持 3 个月。（如图 5-1-1）

而当我们聚焦到数据相关企业，这些企业是否同样受到了疫情的冲击，身在其中的大数据从业者们如何看待疫情对于他们当前的工作与职业发展方向的影响？（如图 5-1-2）

47% 数据团队从业者认为疫情对数据团队工作影响较小

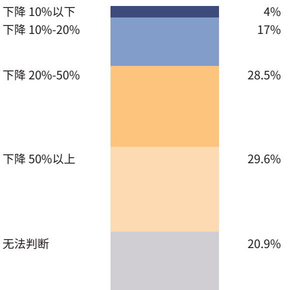
大数据文摘在此前投放的问卷调查中了解到，疫情对于企业中数据团队的业务并未造成很大的影响。在所有受访者中，17.5% 表示自己所供职公司的数据团队完全没有受到影响，29.1% 称几乎无影响，认为产生较大影响的比重为 23.3%。（如图 5-1-3）

六成企业并未裁员，三成受访者的跳槽意愿受到了疫情冲击

从近两年的互联网寒冬，到疫情的直接冲击，裁员问题一直是数据团队从业者的一大隐忧。对此，66.5% 的受访者表示所在企业没有对数据团队进行裁员，大部分企业的运行状况稳定，但也有 15.8% 出现了裁员情况，9.4% 的企业对员工进行减薪，此外，还有 8.3% 的受访者表示，企业同时出现了裁员和减薪的双重问题。（如图 5-1-4）

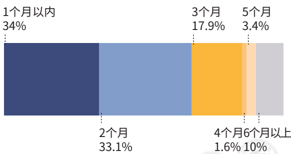
而疫情的发生对于员工的跳槽意愿也产生了一定影响。59.7% 的受访者称自己近期并无跳槽打算。但在有意换工作的四成受访者中，15.8% 在找工作的过程中感到受疫情影响工作机会明显减少，还有 14.4% 的受访者原本有意换工作，但因为疫情影响，暂时打消了跳槽的想法。（如图 5-1-5）

图5-1-1 新冠疫情对中国中小企业营业收入的影响



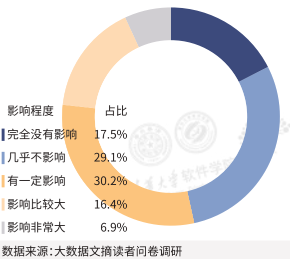
资料来源：艾瑞咨询

图5-1-2 新冠疫情对中国中小企业现金维持时间



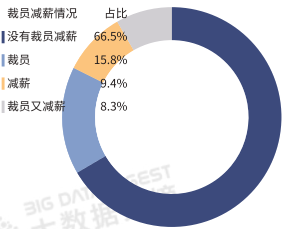
资料来源：艾瑞咨询

图5-1-3 数据团队从业者认为疫情对所在数据团队工作影响程度



数据来源：大数据文摘读者问卷调研

图5-1-4 数据团队从业者反映疫情期间公司数据团队裁员及减薪情况



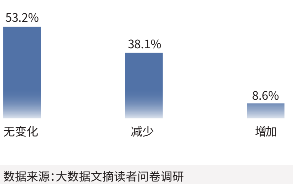
数据来源：大数据文摘读者问卷调研

图5-1-5 数据团队从业者认为疫情对于自己换工作计划的影响



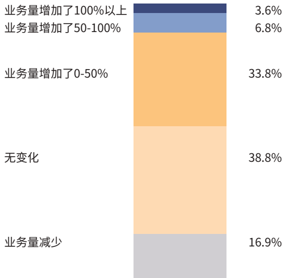
数据来源：大数据文摘读者问卷调研

图5-1-6 数据团队从业者反映疫情期间公司数据团队的招聘需求变化



数据来源：大数据文摘读者问卷调研

图5-1-7 数据团队从业者反映疫情期间公司数据团队业务量增减情况



数据来源：大数据文摘读者问卷调研

超半数企业招聘需求无变化，招聘需求减少的企业占四成

受到疫情影响，部分企业数据团队的招聘需求有所变化。我们发现，有 38.1% 的受访者表示自己所处的数据团队招聘需求有所减少，53.2% 表示没有变化，但仍有 8.6% 的团队反而增加了其招聘需求。（如图 5-1-6）

44% 的数据团队业务量增长，3.6% 团队业务量翻倍

在疫情对于业务量的影响上，38.8% 受访者认为数据团队业务量没有产生变化，16.9% 业务量减小。44% 的受访者则反映数据团队业务量增长，其中 3.6% 团队业务量翻倍。我们可以看到，疫情冲击下，大数据相关业务的价值正在凸显。（如图 5 -1-7）

这一情况也同样在研究组走访多家数据团队的过程中得到了印证。

以联想数据智能业务集团为例，联想数据智能业务集团产品及生态总经理田日辉告诉我们，疫情期间，从用户使用的角度来说，联想数据团队云模式的产品无论从分析还是从报表各方面，基本上没有受到太大的影响。中央的数据团队虽然在家里远程办公，数据用户也在家里办公，但是整个平台运转依然正常。

疫情对团队业务本身供应链会有一些影响，即用户的需求。由于疫情影响，客户对联想的产品，无论是业内产品还是云服务的产品，以及数据智能转型服务这些产品的需求，都是有所提升的。

另一方面，疫情期间，由于多数用户或市场参与者的行为受到了影响，不少企业也因此对产品做出了相应调整，数据团队在这一过程中做出了迅速响应，包括辅助调整产品功能、上线疫情应急方案，团队社会影响力和产品产出能力都得到了相应提升和锻炼。

以领英全球数据科学团队为例，其负责人许亚对大数据文摘表示，领英利用数据优势，在疫情期间实时展现劳动力市场的趋势变化，帮助个人更好地应对当下的不确定性。领英发现，刚入职场的新人会受到疫情更大的冲击，抵御职场不确定性的能力较差，此外，疫情对女性的负面影响也可能大于男性。由此，他们也在业务层面上制定了有针对性的服务措施。

此外，在疫情期间的武汉，滴滴组建起医护保障车队，3 个多月的时间里，共有 300 多名司机加入武汉医护车队，累计为武汉 16 家医院近 2 万名医护人员提供了近 50 万单服务。在全国 15 座城市，共近 16 万名司机自愿报名加入滴滴医护车队，总计服务 37987 名医务工作者。滴滴数据科学团队主要在订单级别从数据角度针对医护人员的用车规律和出行场景进行实时分析，针对不同医护人员的出行特征实现更合理的司乘匹配与路线规划。此外，对出行高峰、出行热区等的预判，也能有效帮助业务团队提前对司机进行调度，更高效地保障医护人员的出行。整个春节，滴滴数据科学团队在相对较小样本的环境下输出了大量分析结果，有效地支撑着医护车队项目决策的快速迭代。

在国内，工信部统一组织下，三家主要电信运营商很快实现了数据整合，为疫情期间的出行和公共健康的防疫管理作出了重要贡献，包括中国移动在内的多家国内主要运营商的大数据团队都紧急参与其中。在中国移动数据团队，整个项目从开启到初版上线其实只花了不到一周，有效的支撑了疫情人群迁徙、行程查询、复工复查分析等各项工作，累计提供疫情防控分析报表上万张。

一支高效运作、抗压的数据团队不仅对于所在公司至关重要，也是整个社会持续、健康发展的生力军。整体来说，疫情对于数据团队的影响并非完全负面。虽然有部分企业出现了裁员或降薪情况，但同样也出现了不少业务量增长的情况，说明数据业务价值在疫情中逐渐凸显。与此同时，疫情对于从业者的职业规划还是产生了一定影响，部分受访者原本的跳槽计划出现了变动。

PART 6
附录

BIG DATA DIGEST 大数据文摘
清华大学软件学院

起底滴滴数据科学团队：面对超复杂线下场景，要数据驱动，但拒绝“唯数据论”



受访嘉宾：

赖春波

滴滴技术副总裁、数据科学与智能部负责人

面对疫情这样的重大社会事件，数据科学团队能做什么？

16 万、37987 名、1500 万公里，这是滴滴数据科学团队在医护车队项目中交出的答卷。

腊月二十九武汉“封城”，大量医护人员出行不便，滴滴随即组建医护保障车队，为医护人员免费提供出行服务，除夕当晚，50 辆车投入运营。如今，即使防疫进入常态化，但每次回想起春节期间的医护车队项目，滴滴数据科学与智能部高级数据科学总监李伟健还是充满了感慨。

3 个多月的时间里，共有 300 多名司机加入武汉医护车队，累计为武汉 16 家医院近 2 万名医护人员提供了近 50 万单服务。而在全国 15 座城市，共近 16 万名司机自愿报名加入滴滴医护车队，总计服务 37987 名医务工作者，行驶总里程超过 1500 万公里。

能够在短时间内组织运营医护车队，除了高效的线下能力，滴滴多年来积累的出行数据和团队用数据解决问题的经验也很关键。李伟健介绍道，医护车队上线初期主要依靠工作人员手动匹配，为了提高发单效率，滴滴紧急为医护人员研发了线上产品，第二天，武汉医护人员就可以在 APP 线上发单。

滴滴数据科学与智能部也在第一时间加入，在订单级别从数据角度针对医护人员的用车规律和出行场景进行实时分析。比如，他们发现，早上七点是医务人员的上下班高峰，很多医生下班后不会回家，而是前往酒店等。

除此之外，对出行高峰、出行热区等的预判，也能有效帮助业务团队提前对司机进行调度，更高效地保障医护人员的出行。整个春节，李伟健都在和同事一起，在相对较小样本的环境下输出了大量分析结果，有效地支撑着医护车队项目决策的快速迭代。

海量数据背后，是滴滴数据科学体系的支持和承接，在大数据文摘采访几位负责人的过程中，隐藏着滴滴的数据基因也逐渐显露出来。

数据体系团队四大模块，助力业务可持续发展

作为一家老牌互联网公司，数据思维一直贯穿着滴滴各项业务的发展。

早在出租车时代的各项服务中，滴滴的数据思维就已经显现，以司乘规模、供需匹配等交易环节为中心，数据分析评估已经覆盖到了体验、司乘生态、城市交通安全等众多场景。

2017 年，滴滴正式组建数据科学部，他们的目标是用数据为滴滴的运营和产品提供洞见，帮助业务在快速的迭代中科学决策，实现“数据驱动”，一方面要重视数据积累，另一方面也要辩证地看待和使用数据。

这对滴滴数据科学团队的后续发展也起到了一定的影响，在数据科学团队内部，奉行着这样一条不成文的规律，先找准业务中最需要利用数据的模块，在这些领域中体现出数据价值后，再将数据决策扩展到其他业务模块上去。

也正是秉承着这样的传统，滴滴的数据科学家们也天然地和业务部门很是亲近。

据滴滴技术副总裁、数据科学与智能部的负责人赖春波介绍道，滴滴的数据体系分为四大模块，大数据架构、数据平台、数据治理、数据科学。在职位划分中，下面三大模块多为工程师、产品经理、数据开发工程师，数据科学分为数据分析师和数据科学家，他们数量最多，以“嵌入式”的方式，分布在不同的业务部门中。其中，数据科学团队，需要在业务形态中实现广泛的运营智能、产品智能和决策智能，助力业务可持续发展。



数据科学：通过系统的数据挖掘和主动深入的业务分析，看清业务发展方向和要素，提出策略建议，帮助业务实现用户价值与商业价值；并通过科学的实验设计和评估，辅助管理层更快更准确地进行业务决策，保证决策质量；

数据治理（DG）：通过系统、管理流程、意识提升等手段，体系化治理全公司数据资产，向前赋能，提高数据使用效率，发挥数据生产力；

数据平台（DP）：通过工具产品，向前提升生产效率、可靠性和可扩展性；

大数据架构（Dinf）：构建稳定可靠、低成本、高性能的大数据基础设施，赋能业务。

2017 年首份《顶级数据团队建设全景报告》调研了解，顶级数据团队一般具有相似的特征：所在组织或机构数据驱动战略明确，数据团队运作高效；高层需要设置清晰的数据团队建设目标并将数据纳入决策流程；数据团队的高效运作则需要优秀的团队领导、合理的组织架构和多样化的人才。

高层中心化的数据指导部门对于一家公司的数据科学团队建设的效用显而易见，包括联想集团、瓜子二手车等公司，都设置有中心化的组织，统一领导公司数据化运营。

滴滴也不例外。根据赖春波介绍，由于滴滴有网约车、车主服务、两轮车、代驾、出租车等多个业务群，滴滴的数据科学家也就很自然地分散在不同的业务部门里。为了能更全面准确赋能业务，滴滴组建了数据科学委员会，增强跨业务数据科学家间的交流和协作，同时对复杂问题进行决策，迭代数据体系建设。

与瓜子二手车的“技术委员会”不同的是，滴滴的数据科学委员会成员占比最多的是数据分析师，他们每季度开会一次，主要针对公司的规划服务和长期定位等进行商讨。

而这些例行会议并不只是技术交流。毕竟除了技术能力和批判性思考的能力外，一个好的数据分析师还需要足够的商业能力、战略视野、影响力、领导力和同理心等素养，每次会议也不可避免地涉及到相关领域的讨论。

“分析师需要把自己脑袋的东西放到别人脑袋，是靠嘴吃饭的。”赖春波笑称。

不过要想真正提升产品、运营和决策的智能化，只靠一张嘴是远远不够的。赖春波介绍，数据科学团队每周会产出四五十份的专题分析研究和每周几千次的实验和评估，这些都随时可能影响到公司决策。前者会呈金字塔式排列，最顶端的体系化和方向性研究是真正实现辅助战略的决策智能，投入的精力也更多；后者主要针对业务或产品的方案进行评估，相对更加自动化和流程化。

构建智慧交通，数据共建共享很关键

如今，滴滴已经成为国内最大的一站式移动出行服务平台，每天处理的数据量高达 4875TB。但滴滴想做的还远不止于此。赖春波说，滴滴希望能帮助构建智慧城市，在交通汽车产业做得更好。

要实现这个目标，仅靠滴滴内部数据是不够的，需要从更大的社会维度进行数据的共建共享。据介绍，滴滴正携手国家预警信息发布中心、各地交通管理部门以及行业合作伙伴，进一步丰富平台天气特征、路网信息，积极鼓励司机和乘客进行交通上报，加强数据的完善。

也正是得益于与外界的众多合作，滴滴在二十多个城市基于平台车辆数据及城市交警卡口、地磁等多元数据，落地了包括智慧信号灯、智慧交通诱导屏、交通信息系统等智慧交通项目。不仅如此，滴滴还向学界免费开放脱敏后的出行场景数据，助力学界更好地进行前沿探索。

从 2017 年接入 ofo，2018 年正式托管小蓝单车，到上线自有品牌青桔，伴随着用户骑行数据的不断完善，滴滴还与公交集团开展定制公交、实时公交等合作，用户出行生态就能够在多维度进行描绘。

但得到数据还只是第一步，全球 AI 咨询公司 Pactera Edge 副总裁 Rajeev Sharma 在提到数据系统时

曾说过这么一句话，“你面对的，是一个有生命的、会呼吸的系统”。

数据在体系内的应用则更加重要。对供给和需求进行有效预测，并提前进行调度，是提升网约车效率和服务体验的关键。温度、降水、司乘活动，以及是否是工作日等都会对供需情况产生影响，基于海量数据和机器学习等算法，滴滴也能模拟未来供需和历史供需、司乘补贴、城市、天气、节日关系，进而对未来的供需情况进行预测，进而更加精细化地提前进行调度，进一步提升成交率和司乘两端的出行体验。

而在安全层面，数据分析也在发挥巨大的价值。 由于疲劳驾驶危害极大，通过大量的分析，自去年 6 月起，滴滴在《道路交通安全法实施条例》要求的基础上还上线了防疲劳驾驶的长时策略，司机达到一定计费时长后休息 6 小时才能上线；之后又进一步在车载录像设备中设计了疲劳驾驶预警系统，能在设备中自动进行图像处理与分析，检测司机的疲劳特征，在司机进入疲劳状态前语音提醒司机注意行车通风、及时休息。

相关分析还显示，不仅仅是在夜间，在凌晨或午后一些其他时段，司机重度疲劳的概率也会更高，为此，滴滴扩大了易疲劳时间段的覆盖范围，开始在易疲劳时段对全量司机进行实时的加强提醒，以进一步帮助提升驾驶安全。

依赖数据，但不“唯数据论”

这些工作能够顺利完成，与滴滴内部已形成共识的数据文化思维有密切的关系。实际上，广受社会关注的安全事件发生后，滴滴还开始在网约车引入“安全派单”，这对数据模型的要求也变得更高。派单系统需要分析评估发单场景可能存在的风险，充分考虑乘客的性别、出行习惯、订单时间、订单距离等订单特征和司机的驾驶习惯、历史订单信息、投诉记录等服务质量分级，

在此前全局派单的基础上来进一步综合分析司机和乘客是否合适出行。

赖春波也坦言，安全派单一定程度上可能也会伤害乘客体验，出现女乘客深夜有时更难打到车的情况，策略还需要不断调整优化。

也正是有了这样的案例，对于如今的滴滴来说，数据面临的问题不在于对“数据驱动”本身的质疑上，而在于当评估业务中一个不可量化的任务时，如何把握好“度”。

交通行业是一个系统工程，很多环节面临的问题都没有也没有任何一个完美的模型可以解决。一边要千方百计从技术上去做更好的算法，但同时也需要企业更多考虑现实和线下场景的复杂性。

滴滴数据治理和数据平台部负责人王勇总结称，“要依赖数据做决策，但不能只依赖数据做决策。”

在他看来，和电商业务不同，滴滴的双边交易市场是实时地一对一撮合交易，是处于更加零和博弈的状态。同时，移动出行的行业渗透率有限，滴滴兼具线上和线下特征，线下数据显得十分重要，但行业内专家相对有限，这就需要大家持续探索利用数据进行不断地试错迭代，沉淀方法论和战略视野。

李伟健也表示，正是这些让滴滴场景里的实验和评估等数据科学问题变得更加独特和有价值。比如当滴滴在乘客端做了实验优化后，却发现因为司机端运力被抢夺导致结果不理想。“这里面有很多有挑战性的问题，我们也非常欢迎优秀的统计人才加入。”

据介绍，为了更好地将数据文化贯彻到整个公司内部，滴滴数据体系还面向全员开设了以提升数据能力、培养数据思维为目标的能力提升课程，并结合技术分享、训练营等多样化方式展开交流。

针对数据科学家，则会在专业技能之外会更侧重培养商业理解能力、洞察力和影响力，以进一步提升通用素质和专业技能。在学习互动过程中，加强团队信任，通过切实案例理解数据驱动，这对每个岗位来说都是必要的。

在数据应用落地上，为了让数据、业务、工程三方合作更和谐，滴滴成立了 Data Business Partner 团队，既要强调客户意识导向的文化，也要强调价值牵引技术驱动的背景，很多时候只依靠任何一方都是不行的。

在谈到对数据人才的期待时，滴滴表示，他们会重点关注以下五个方面：

- 用户导向思维，数据发挥价值要跟创造用户价值联系在一起；
- 认识数据的边界和局限性；
- 要有同理心，数据方虽不直接负责业务，但需要了解彼此目标；
- 布道者角色，让数据文化落地，影响更多人；
- 敬畏数据安全和隐私。

“数据科学家”或许不再性感，但“数据团队”的产业化才刚开始 | 专访领英全球数据科学团队负责人



受访嘉宾：

许亚

领英全球数据科学团队负责人

定下“顶级数据科学团队”这个研究话题时，我们第一时间想到了领英（LinkedIn）。

2008 年，正是在这家公司，DJ Patil 建立了全球首个真正意义上的“数据科学团队”，并开始用“数据科学家”（Data Scientist）这个词来描述这些 Data man 们的工作性质。

在这之后，“数据科学家”开始被誉为 21 世纪最性感的工作，也成为全球技术精英们近年来最理想的职位之一。

尽管已经过去了十多年，但当我们请领英全球数据科学团队负责人许亚给数据科学团队下个定义时，她还是表示，这不容易。

的确，尽管数据科学在学术领域的概念 50 多年前就有了，但作为职业，相比业内更多成熟的团队和路径，这依然是个相对很新的概念。

不同公司和团队领导人对于“数据科学团队”的定义范畴大相径庭：

从时间维度来看，当年研发出 Hadoop、Kafka 的人会称自己是数据科学家，但是现在这些大数据底层技术都变成了偏基础设施的内容，在狭义概念上，已经不再属于数据科学团队的范围；

随着这个领域囊括的范围越来越广泛，数据对于每家公司的重要性也都只增不减，数据科学的“嵌入”性越来越高、边界也越来越模糊。

尽管如此，谈及领英这些年“数据科学团队”的定位和建设，许亚依然有自己非常清晰的思考。

“对于领英来说，数据科学团队的整体趋势更加走向专业化，他们的职责不再是建立数据基础设施或平台，而是怎样去使用数据科学和工程来最大化数据的价值。”

这是许亚对数据科学团队任务的要求。

那么到底如何让数据的价值最大化呢？从团队运作方式、商业影响力设定和社会责任等角度，许亚给出了领英的答案。

“嵌入式工作，中心化管理”，数据科学团队更加“专业化”、“工程化”

和多数互联网公司一样，领英的数据科学团队规模也在近几年飞速增长。许亚表示，仅是近两年来，领英的数据团队扩张了近一倍，从 150 人增加到目前的 300 多人。

许亚提到的数据团队是指领英中心化的数据科学部门。如果用一句话来概括领英的中心数据科学团队的运作方式，那就是“嵌入式工作，中心化管理”。

和国内不少互联网公司将数据分析师归属于业务 BU、向业务主管汇报不同，领英的数据科学团队成员由许亚的中心部门统筹。虽然在项目工作上，数据科学家们依然会在工位分布和职能上与业务部门紧密联系，但是从职级从属上，都直接向许亚汇报，不同领域的数据科学家在工作中会有交集，还会一起开会。

其实领英的数据科学团队的设置也不是一开始就如此，随着领英数据科学团队定位的变化，数据科学团队也从最初的产品组，移到了现在的工程大组。

值得一提的是，目前领英的数据科学和人工智能团队都在同一个大组里，许亚表示，数据团队和人工智能 / 工程团队是紧密相连的。

这也从一个侧面说明，随着对数据科学团队的需求逐渐增大，数据团队的工作会越来越“工程化”。跑的数据会越来越多，对工程团队的需求也会越来越大，需要对工程团队越来越多的要求和技术定位。

近年来，各大公司越来越意识到数据的重要性，已有的数据科学涉入领域在进一步扩张。数据团队之前最常被用到的部门是市场和产品，但是基于领英本身的数据基因，近几年的一些产品也对之前没有用到数据的地方做了数据驱动的尝试。

例如，与架构工程部门合作的数据团队会去衡量工程架构的建设是否有效率：每年跑大数据的硬件设备花费很高，怎么样在时间上做规划，让硬件 /GPU 等更有效的发挥价值。

在人员构成上，和十年前相比，领英的数据团队也更加专业化了，底层架构人员也从数据科学团队分离了出来。

目前领英的数据科学团队也根据员工不同的专业领域设立了三个工作方向：

- 工程专家：可以很有效的建立起数据管道 (data pipeline) 和数据流 (data flow)；
- 算法专家：在预测、算法领域的技术咖；
- 业务专家：有很强的业务属性，将数据见解和公司战略结合起来；

由于工作侧重不同，在管理的过程中也会有意的区分这三类数据科学家，并且保持各类员工的竞争力。

许亚提到，她的团队内部更多是自下而上的工作文化。她不会给团队指派任务，因为每个组会自觉的告诉许亚他们想达到什么样的目标。对于一些大的项目，一般需要跨部门合作，各部门的领导达成共识，分配资源来一起实现这个目标，是自上而下和自下而上的结合。

三大 KPI 指标，量化数据团队工作

相对复杂的构成和与业务团队的紧密性，给数据团队设定商业影响力和发展路线不是一件容易的事。

许亚表示，两年前她接手领英数据团队后做的第一件事就是拟定了团队成功的三要素。虽然数据团队的价值有时候很难量化，但是有三个指标可以作为探讨的基础。在数据团队内部不同组可能会有不同的侧重，但对大部分组来说这三个因素都很重要。

数据易得性和工作效率

数据易得性，指的是当外界需要数据的时候，获得这些数据的难易程度；工作效率，指的是一个人的工作是否可以提升整个团队的工作效率。

许亚表示，数据科学家之前被人诟病过于追求新鲜感，喜欢挑战高难度问题，但做完 MVP (Minimum Viable Product) 后没有维护迭代的习惯，永远都在追逐下一个新难题。数据团队拥有许多数据资源，比如原始数据，指标数据，数据模型，数据可视化。

当外界对这些资源有需要的时候，如何能够保证这些需求能够随时被满足？软件开发有一系列衡量数据获取难易程度的指标，比如 SLA(Service-Level Agreement) 的达标率就是一个很好的量化指标。

有些数据科学家做了一个很不错的分析，但是不太关心怎么把这个分析过程自动化，所以每次有人提需求的时候就需要有人再手动跑一次模型，其实都是重复劳动，不同的人在做相同的重复劳动。如果这个分析实现了自动化，大家都可以享用，其他人就不需要花太多时间精力在这个模型上，整个数据科学团队的集体工作效率都提高了。

以前许亚的团队也缺少这种分析自动化产品化的意识，所以她把这个设置为成功三要素之一，强调这种意识的重要性。

战略化思维

战略化思维，指的是数据分析结果对公司重要战略性决策是否有指导作用。

许亚的数据团队和公司很多高层会打交道，因为他们团队有一个很重要的职责就是通过数据来确保公司重要决策的大方向是准确的。比如他们需要了解用户在疫情期间是如何使用领英服务，如何通过领英的产品获取价值的。

许亚认为在疫情后，用户的行为多少会发生一些不可逆转的改变，数据可以帮助团队更好地去学习用户行为变化，从而在战略上指引公司对哪些领域进行重点投资。不管是产品开发还是市场战略的决定，都需要依靠数据。

直接商业影响力

直接商业影响力 (Direct Business Impact)，指的是工作成果对公司商业目标的直接影响力。

每个部门的工作开展是和公司要实现的大目标息息相关的，领英有公司层面的四个核心指标，数据部门在计划工作的时候，需要考虑如何对公司的商业目标产生积极影响。

AB Test：用数据来证明一切

我们都知道，企业在做产品 / 功能测试时一般都会用到 A/B Test，即分为两组用户，一组对照组，一组实验组。对照组采用已有的产品或功能，实验组采用新功能。要做的是找到他们的不同反应，并以此确定哪个版本更好。

A/B Test 能对大范围的事情进行测试，例如亚马逊对个性化推荐进行 A/B Test 后，发现个推能显著提升收益；谷歌在对搜索广告进行排名时也用到了 A/B Test。

那么对于领英来说，A/B Test 在领英的产品设计中又扮演着什么角色呢？如何影响产品决策呢？

可以这样说，基本上我们在领英网站上能感知到的更新，领英团队都会做 A/B Test，有些是前端的改变，有些是后端系统的调整。当你打开领英 APP，从搜索栏，搜索引擎算法，底部导航，到页面文字大小，这些都是经过 A/B Test 的。

领英的产品文化以用户为主导，领英自己不会去假设用户喜好，一切都通过数据来说话，而不是靠谁的直觉。除了看得到的东西，后端用户看不到的，领英也会进行 A/B Test。比如打开 APP 要加载内容，需要从后端系统里获取数据，每次获取 20 条数据还是 100 条数据，这个决策就涉及到平衡与取舍，获取数据越多，页面加载时间越长；获取数据越少，用户浏览的时候就需要频繁刷新。所以到底一次获取多少数据，领英还是通过 A/B Test 来决定。

还有一个简单的例子，当领英对一个数据中心的开关做决定时也依靠 A/B Test，比如一个用户发起数据请求，这个请求该发送到哪个数据中心来处理呢？这种情况下用户到数据中心的距离就是一个很重要的考虑因素，最终领英会通过做 A/B Test 来选择最优化的基础设施方案。

虽然数据团队是 A/B Test 方面的专家，在这方面更有经验，但因为领英有非常完备的 A/B Test 平台，可以解决大部分实验需求，包括实验设计、实施和分析，所以数据团队不需要介入到每个 A/B Test。

这对推广实验文化和数据文化很有帮助，因为大家都可以去做实验，享受数据和实验带来的好处。领英内部每天大概有 100 个新实验在进行，数据团队无法关注每个实验，但是会集中关注一些重要的实验，深入参与到研究和分析工作中。

在领英以数据为主导的文化浸染下，长远来看所有人都受益于这样科学的决策机制。也因为有 A/B Test 的文化，所以可以跳过争论，直接做个 A/B Test 就见分晓了。整个过程简单公正，方案落选的组也可以通过这个机会学习到一些关于用户的新知识。

A/B Test 提倡数学引导的创新，这种创新不取决于谁的职位更高，因此任何团队都可以放心大胆的去做测试来发掘新点子。

领英作为一个社交平台的社会责任：给每个人公平的机会

在许亚看来，维护公平是一个很有挑战的课题，因为你很难明确定义公平。

“当我们在说公平的时候，我们在说公平的机会？公平的结果？还是公平的待遇？我之前看过一个有意思的问题，给三个不同高矮的人提供凳子，在公平原则下，你该给他们提供同样高度的凳子？还是提供不同高度的凳子让他们坐上去之后一样高呢？我很难说这个问题有一个绝对正确的答案。”

领英对公平的定义是，拥有同等才能的两个人，应该获得同等的职业机会。而不受到种族或者自身人脉的影响。过去两年时间领英做了很多努力来解决公平问题，取得了不错的成果。

首先，领英很重视可量化、可测量的指标，因为如果一个问题没有被数据抓取到，就很难注意到。

例如，每次领英发布新产品，都需要通过量化的指标来测量这个新产品对用户带来影响是否公平。一开始领英的测量指标比较粗线条，他们会看这个产品平均下来对用户是否有积极影响，但如果细看数据，有可能这个产品只对一部分人有益，但会损害另一部分人的利益。因此，后来领英采用了一个指数来衡量是否在一个群体内无意间引入了不公平因素，也就是对每个新产品，领英想知道其带来的提升是否是公平的。

其次，领英关注现有平台上是否存在公平问题的盲点。

例如一个以男性为主体的数据集，训练出来的模型就更倾向于男性，这是一个隐蔽的不公平点。很多猎头和 HR 用领英产品来招人，如果算法推荐的候选人都是男性，女性就失去了公平的竞争机会。

大概一年前左右领英推出了一个代表性指数来衡量推荐结果对整体数据集的代表性。比如所有可能候选人的男女比例是 1:1，那领英给猎头推送的前 100 位候选人的男女比例也应该是 1:1。有了这些量化指标，领英可以更好地规范和规避不公平的举措。

许亚还给我们举了一个例子。之前领英有一个内推功能，当某个人想申请 Google 的工作，会收到提示说我的一位好友在 Google 工作，我可以找他要个内推。

上线初期，领英内部对这个新功能很满意，因为可以帮助那些有广泛人脉资源的人更快找到工作，后来领英意识到这个功能会让那些没有人脉资源的人更难找到工作，所以就关闭了这个功能。取而代之的是领英推出了一个新工作快速提示功能，一个新职位刚发布出来，领英会立刻给所有对此类职位感兴趣的用户推送提示。这个功能不仅能帮助所有用户更快找到工作，对那些关系少的人尤其有帮助，因为他们的消息相对更闭塞一点，所以这个功能能让更多的人受益。

最近领英也开源了这套技术，希望能助力其他公司去构建一个更公平的社会环境。

随着近年来数据泄漏事件频频爆发，数据隐私和安全问题被推上了风口浪尖。许亚也跟大数据文摘聊了聊领英在保护用户的数据隐私方面都做了什么。

领英全球有超过 6.9 亿用户和 5000 万家企业，领英的愿景是为全球劳动力市场中的每一位创造经济机会，通过将所有在领英平台发生的行为数据可视化，进而打造全球“经济图谱”。因此用户数据对领英至关重要，如果没有用户的信任，领英就没有办法去实现他们的愿景和使命。

所以在 GDPR 这些开始之前，领英在保护用户隐私上已经有了很多投资。许亚提到，除了实现规定里的要求，领英也用一些很前沿的技术去确保不泄露隐私，比如现在认为是数据隐私保护的“Gold Standard”——差分隐私（Differential Privacy）。

大家经常说到保护隐私，比如说把一些个人信息隐去了，其他人看不见，我就没有隐私泄露了，其实不是这样的。

差分隐私只是一种保证。假设你的信息在一堆数据里面，如果把这些信息删掉，再运行同样的一些算法，从数据当中得到的两个的结果都是一样的。相当于你的数据在或者不在这个数据库里面，最后对于得到的信息没有影响。这样用户就不需要担心他们的数据隐私被泄露。

领英三年前就开始针对数据隐私问题进行一些重要的研究，同时也有一些比较成功的应用，例如最近一个针对广告商的产品，客户想要用领英的 API 去获得一些信息，比如用户互动量前十的文章，像这样一些集合的信息，领英也用差分隐私去确保用户的信息不泄露。

最后，从整个公司文化上面来说，许亚透露，除了去实现数据保护条例的一些要求，领英也用到了一些很前端的技术，来确保用户的隐私不被泄露。另外，领英也十分重视在数据分享方面的问题，并表示会对此加强技术防护。

采访过程中，许亚多次提到领英的社会责任。今年，一场突如其来的疫情，全球的劳动力市场都受到了不同程度的影响，不论是就业还是工作方式都迎来了一种新常态。领英利用数据优势，实时展现劳动力市场的趋势变化，帮助个人更好地应对当下的不确定性。在分析数据时，领英还发现不同分组内的用户受到的影响程度不一样，比如刚入职场的新人会受到更大的冲击，疫情对女性的负面影响可能大于男性。

通过数据观察到这些问题后，领英数据科学团队和业务部门迅速沟通，快速响应，针对各个市场及时提供了一系列有针对性的服务来帮助这些人，让每个人都能在自己能力范围内获得平等的工作机会。

“这是领英作为一个职场社交平台的社会责任。”

数据团队「隐形守护者」！从被动应对到资源输出，腾讯安全 20 年成长记



受访嘉宾：

黎巍

腾讯安全副总裁

根据最新发布的《IDC 全球网络安全支出指南》预测，2020 年全球网络安全总投资将达到 1202.8 亿美元，较 2019 年同比增长 10.1%，而中国网络安全市场总体支出将达到 87.5 亿美元，较 2019 年同比增长 24%。与全球相比，中国网络安全市场近几年在国家政策法规、数字经济、威胁态势等多方需求驱动下，整体市场规模发展快速。

但同时，我们也应清醒地意识到，虽然中国网络安全投入在整体 IT 投入中的占比有所提升，但相对于全球平均水平还存在一定差距。国内的大量企业要么处于被动防御状态，要么缺乏自己的专业安全团队。

随着产业互联网的发展，尤其是随着年初疫情的爆发，加快了实体业务数字化上云的步伐，随之而来的，企业的安全防护特别是数据安全的防护也在面临新的挑战。

这也让企业的安全团队承担起“隐形守护者”的角色：这不是一个高调光鲜的工程，需要日积月累的持续建设，甚至你在日常工作中可能都感受不到它的存在；但是一旦有重大事故发生，就可能给公司造成巨大不可挽回的损失。在高速发展的产业互联网时代，大量新技术的出现让安全问题已经不是单点领域问题，而是一个系统工程，数据安全的应用范围早已超出了数据本身的安全，还涉及到整个安全体系。



作为科技领域的领军企业，腾讯安全团队在二十年的运营过程中，为云上安全积累了宝贵的经验。六月初，我们跟腾讯安全副总裁黎巍聊了聊腾讯安全团队的建设问题：作为腾讯整个数据系统“保护者”，安全团队如何一步步发展起来，又是如何将这种能力输出给到行业的。

要做好内部建设，也要走出去：安全团队建设的三个阶段

腾讯自身的安全建设，在二十年的历程中，经历了三个阶段。

第一个阶段是启蒙阶段，成立初期的腾讯和其他公司一样，安全建设以防御和对抗黑客入侵为主。

当然，要建立自己的安全团队不是那么简单的事，制定安全规范、构建安全体系，这些都是必要的。随着后期腾讯业务不断扩展，团队发现很多安全问题具有共性，如果只是一味被动应对，不仅会陷入“持久战”，团队也很容易进入疲态。

所以 10 年前，腾讯安全开始主动做一些安全的运营和建设，也正是这个时期安全团队的建设进入了第二个阶段，即把安全体系化和产品化，进行主动地运营。这个阶段，安全团队就总体目标达成一致——保证核心资产数据不会被窃取和丢失。这也是二十多年的发展中腾讯安全一直在践行的理念。

到第三个阶段，腾讯自身的安全生态已经做得比较系统了，但是放眼国内，还有不少企业在安全方面处于非常原始的阶段，这就触发了腾讯安全想要走出去，把 20 多年的安

全经验和能力资源输出到整个产业中，帮助产业数字化转型。

黎巍坦言道，腾讯安全希望未来不只是为企业提供产品或解决方案，还能够为企业转型打造合适的安全战略观，更多维度的能帮助企业解决安全问题。

“超过 90% 的企业一定是在云上更安全”

但是也正如黎巍所说，“数据不是静态的，是动态的”，数据安全问题本身就是一个复杂的系统，这也是企业不愿意主动处理安全问题的原因之一。

疫情之下这个趋势更为明显，传统行业都在经历着数字化转型，原来的静态资产和业务逐渐数字化，进而用于服务客户，这是典型的数据流动问题。在这个高速流动的数据体系里，想要更好地实现防御安全，系统管理的视角必不可少。

去年，在腾讯的全球数字生态大会上，黎巍曾经表示，“90% 的企业一定是在云上更安全”。经过了疫情考验，黎巍告诉我们，从目前的实际情况来看，这个数字肯定要超过 90%。

受疫情的持续影响，以线下业务为主的中小企业收入明显下滑，线下模式受创后，众多企业纷纷转型线上。

但是面对迫在眉睫的生存问题，这些临时转型线上的企业往往会忽视重要的安全问题，或者因为资金和技术原因，暂时把安全事宜放在一边。这种情况下，从云上获取安全保护对企业来说是一个很好的选择。由于云上已经构建好了完善的安全体系，企业只需在这个基础上投放业务即可，这与从零开始建设安全团队和安全防御产品相比要容易得多。

新技术触发新的安全危机：伴生型的安全问题，需要前瞻性的安全意识与体系化建设

近年来，随着物联网、AI 大数据、5G 等一众新技术的出现，安全问题变得更为复杂。

这种技术的发展一定会带来新的变革，因为安全是伴生型的，一定会伴随着新的业态，演变出新的安全状态。因此安全团队的主要挑战在于，快速适应新兴的业态。

根据多年累积的经验下来，腾讯安全也总结出了一些不变的原则，对于此，黎巍分享道，首先是对安全问题的零容忍，只要出现安全问题，就当做头等大事应对，这是很重要的第一点。

另外，腾讯始终坚持践行最小安全权限原则，用现在比较流行的说法就是零信任。其实在 10 年前，腾讯安全就已经进行了初步贯彻，具体来说，就是对企业网络内外的任何人、设备和系统，都基于不信任原则进行系统建设。

最后就是安全能力的积累，腾讯有二十年黑灰产对抗的经验，这些经验的积累再加上威胁情报的分析，

可以根据不同的业务、不同的业态进行快速地研究，构建产品和解决方案。

但是就国内整体环境而言，在安全方面的投入和意识都存在较大的缺陷。对此黎巍总结道，首先多数的企业缺乏组织安全团队的经验，其次，企业大多处于被动响应状态，不出问题就当成没问题，第三，企业的安全意识不够，与专业安全公司的互动也不强。

至于在安全方面国内企业能否实现“弯道超车”？黎巍坦言，与其期待超车，不如持续扎实的进行安全建设，毕竟底盘不稳的话，弯道翻车的情况更多。等底座扎实后，自然就有能力应对各种威胁了。

同时，国外的不少企业和团队安全体系相较而言都要系统得多，可以进行适当地参考学习。

对于如何将安全意识下沉到公司内部，黎巍表示，这一定是自上而下的过程。近几年从全球来看，安全方面的立法立规不断增强，比如欧洲 GDPR、美国 CCPA，中国也出台了一系列数据安全法规。这其实就是从顶层驱动告诉大家，安全的重要性。

二十年前，腾讯就已经重视起了安全问题，如今腾讯内部有两大安全团队，安全平台部与企业 IT 部，这两个部门会根据国内外的法律法规的制度层面进行相应的内部建设。除此之外，运营和产品也都遵循着严格的安全体系和制度规范，比如在产品研发阶段，甚至会细化到某些高危函数的使用。

用 AI 预警新型威胁，人机协同仍是安全团队主要工作方式

去年，腾讯安全推出 AI 预警技术，进一步地保障线上安全。

黎巍介绍道，其实在很多年前，AI 就已经在安全领域做出了不小贡献。就国内互联网行业传统的安全观念来看，会更偏向硬件，而近年来随着物联网、云的出现，安全问题就

变得复杂起来。对企业而言，面对越来越多的新型未知的威胁，需要更加智能高效的系统来进行预测和判断。

就腾讯的所有安全产品而言，它都有一个基座，AI 就扮演了这样的一个角色，在这个基础上进行协同，深入到各种产品中，再进行更广泛地应用。

在腾讯安全内部 AI 的主要运用有两点，一个是对一些新型未知的威胁的预测和预判，还有就是协同的联动防御。

举例来说，腾讯安全有终端也有云，有时候在终端上发现了一些新型威胁，这就需要一个全链条的联动，把这种威胁贯通到云上。反过来也同样成立，这其实算是一种产品跨终端的联动，也是 AI 在基础驱动上的价值体现。

不管是对新型未知威胁的预测，还是协同联动防御，人机协同仍然是腾讯安全在提供服务时主要的工作方式，黎巍介绍道，人类和机器其实是一个相辅相成的过程。

腾讯安全每天的业务流量是非常庞大的，在这种海量的业务流量的情况下，要找到新型未知的威胁，本身就是非常大的挑战，这其中我们就能通过一些算法、AI 技术，更好地进行分析和预警。

目前 AI 和大数据处理技术已经广泛应用于安全领域，不过黎巍认为 AI 仍然处于初级阶段，一些高级威胁分析还是离不开专业安全工程师的深度参与，但面对复杂的企业数字化业态以及云时代的海量数据安全挑战，他相信基于 AI 的安全技术演进将重塑安全产业，也将助力企业更加高效的应对数字化转型过程中伴生的各类安全威胁。

近 10 年数据智能团队建设，联想总结了由内而外的发展经验 | 专访联想集团副总裁田日辉



受访嘉宾：

田日辉

联想数据智能业务集团产品及生态总经理

去年6月，联想集团公开宣布，成立数据智能业务集团（Data Intelligence Business Group，DIBG），由蓝烨担任高级副总裁，直接向杨元庆汇报。同时，联想集团副总裁田日辉负责数据智能业务集团的产品和生态业务，汇报给蓝烨。

在当时写给数据智能业务集团的内部信中，联想集团董事长兼首席执行官杨元庆表示，“在大数据积累的基础上成立联想数据智能业务集团，是为了加速智能化变革，是实施联想行业智能（Smart Verticals）战略的重要举措。”

这也是一直被外界誉为信息化标杆企业的联想，向数据智能领域延伸业务的一次重要的商业布局。

其实联想内部广义上的信息化早在 2000 年左右就开启了。经过 20 多年的信息化实践和积累，联想目前的大型主系统有上千套，整个公司服务器加起来有上万台，数据链量级已经达到十几个 PB。

同时，联想在全球都运营着庞大的上下游生态，在全球两百多个国家和地区同步进行用户服务与体系构建，目前有 50 多家上游企业，2000 多家下游渠道企业。

经过一年的发展，联想数据智能业务集团发展如何？联想集团的数据平台构建与数据团队建设又有怎样的进展？针对以上问题，上个月，我们对联想数据智能业务集团产品及生态总经理田日辉先生进行了一次专访。

三阶段数据团队建设，打造全价值链数据智能平台

在成立数据智能业务集团之前，联想内部的数据团队其实已经初具规模，团队建设也不是一蹴而就的。

自 2011 年联想开始应用大数据至今，联想的数据平台建设主要可以划分成三个阶段：

- 第一阶段：企业内部先进数据应用建设；
- 第二阶段：在内部和外部构建平台；
- 第三阶段：构建上下游企业的数据智能生态。

2011 年到 2015 年是联想数字团队建设的起步阶段。当时的数据团队还主要是服务联想内部拓展数据的应用，因此数据团队的规模很小，只有十几个人，四年后慢慢扩展到一百多人。

田日辉告诉我们，当时团队研发中国第一款安卓系统手机乐 phone，通过应用商店、SBK 模拟器、开发环境与数据分析工具等来构建一套大数据体系，进而帮助应用商店开发者收集相关使用数据。联想的数据团队最早就是从此入手，帮助应用商店的开发者分析月活、日活与产品质量等数据，持续不断优化这些智能应用。

田日辉表示，在初期内部硬件的生产上，联想就对于信息化比较重视，是国内首批使用 ERP 的企业之一，因为自 2004 年收购 IBM 的 PC 业务后，全球化运作对信息系统的要求较高，IBM 原本的很多系统被逐渐废掉，联想也构建了自有系统，提升效率。

之后随着这些应用的深入，大数据不仅能在应用层对产品提供优化，企业内部大量的运营数据，对企业产品研发、供应链管理、市场营销还有很多影响，使得数据应用可以扩展到整个延产供销服务全价值链。

2016 年到 2019 年下半年，联想开始进入了数据团队建设的第二阶段，开始把数据应用大规模进行平台化和推广，并开始将能力对外输出，拓展给外部用户，并在公司内部构建平台。

这一阶段，联想开始有计划的建设全公司“人人都能使用”的数据平台，“这个过程中，公司内部核心策略就是人人都是分析师。”

田日辉提到，数据团队负责建平台，把核心数据和公用数据帮用户整理好；业务人员（IT 部门、业务部门等）完成平台分析工作，大量人员参与进来，使得业务经验能够与数据分析方法更好结合起来，以自服

务式应用致力于解决数据科学家缺失的难题。数据团队开发平台性产品之后，在公司内部先部署与使用，比较成熟之后，再服务外部客户（汽车、石化、钢铁等企业）。外部客户因为与内部的应用模式有所不同，有很多创新的需求和应用点，通过反馈去优化产品，增加内部应用平台的功能特性，并建立内外部互动模式。

而到了现在，联想数据团队建设已经进入了构建生态化阶段。联想所处的行业生态链中，上下游企业都处于数据智能转型时期，联想数据团队现在已经构建了上下游合作伙伴都能利用的智能化应用平台。平台除了提供算法工具之外，还预置了多年积累的分析数据模型，包括预测、仓储优化、用户画像、精准营销等，使合作伙伴能够很快地使用，把数据与业务模型对接，实现业务价值闭环。上下游企业可以选择用联想平台去做智能化软件，也可以使用联想的产品构建私有平台。下游大量中小型制造业企业受限于信息化成本，在数字化转型过程面临很多挑战，建立生态系统可以使其直接使用经过实践验证的平台。

联想希望在对外服务的同时，也能够建立起生态系统，服务更多的尤其是中小企业客户，推动中国智能制造快速转型。

内外部数据治理结合，优化数据平台结构

细数数据团队建设的三个阶段，田日辉对第二个阶段的印象最为深刻。

和很多业务部门较多的集团型企业一样，业务规模如此庞大的联想也面临着数据分散在不同的业务系统中，难以整合的痛点。数据团队建设在初具规模之后的最重要任务，就是建立起一个更完善的企业级数据分析平台，把这些内部分散的数据以集中的方式进行整合管理并科学地利用起来。

正如上文提到，联想注重业务与数据团队的紧密结合，内部的数据团队与不同业务部门分工合作。因此从 2016 年开始，联想就开始把联想几十年信息化中的大小数据系统整合起来，形成企业整体数据湖，并构建统一的数据模型。

数据团队在中央提供分析算法工具，积累推广模型，提供复杂建模过程的二线帮助并举办培训活动，辅助业务部门采取自助式的服务。

田日辉给我们举了个销量预测的例子，说明联想数据智能团队内部是如何使用内部流程化工具为业务部门提供辅助的。

联想生产销售各种复杂的设备，因此销量预测是多层次的，总销量预测会分不同地区和不同产品线。在不断发展中，数据团队把预测模型放到平台上，通过几轮配型后，进行模型积累。平台本身提供很多分析和算法工具，使业务人员运用不同的数据级，使用自动化机器学习工具测试不同的算法，并给出最优结果，同时根据业务实践来判断哪个参数和配置最符合要求。

由于相关数据表极其庞大，可能会存在一些数据冲突，因此公司级大平台可以进行统一数据治理，让所有人的分析工作达到比较好的效果。

要做一个有效的预测或者优化，数据链，尤其数据的广度是非常关键的，因此联想也引入了很多外部数据。在内部平台建设的同时，联想接入大量外部数据进行合作。合作的供应商、市场预测部门、客户满意度调查，各种产品质量反馈等数据都会作为外部数据汇总进来。

内部的数据平台建设逐渐成熟后，联想的数据团队也开始将数据能力输出给提供商与服务提供商，更需要深入理解客户业务和机理模型。

对于外部行业客户，联想内部的数据科学家在专业知识理解方面相对薄弱。在进入行业初期，团队与客户的行业专家一起做项目，客户对企业的数据积累情况与行业的机理模型更清楚，而团队对数据与算法比较清楚。渐渐，客户本身会具备数据使用能力，团队也会积累一些所谓的行业专家，进而把应用模型带给其他客户。由于很多案例和应用框架可以复用，团队也一直在尝试加强对行业的理解，建立一些行业专家人才队伍。

因此，田日辉对于意向进入数据科学领域的高校学生，也提出了一些行业知识的期待。

“掌握新技术有较好的基础，且自学能力与使用能力强。但是应该更多理解企业的运营模式，业务需求和机理模型，多参加一些真正与实战相结合的活动，或到企业里面参与一些实际的工作与项目。”

明确团队绩效指标，“不是一件难事”

很多团队关心数据团队本身的商业影响力及其产生的价值，相比一些传统的部门，数据团队的价值比较难估计与量化。

但在联想内部，估算产出投入与价

值却并不是一个很困难的事情。田日辉提到，由于在第二阶段建设了全集团统一的数据平台，只需要了解公司内外对于数据平台的调用有多少，并且在平台内的操作帮助业务部门创造了多少价值即可。

具体来说，团队关键 KPI 有用户数，产生的价值量等。由于最终的应用由业务部门完成，完成这个项目后，它的质量提升了多少，它的预测精准度提高了多少，意味着多少业务价值，都是很明显的。

到了第三阶段，对于上下游企业来说，外部企业客户最直接的价值就是其购买产品与服务花费的数额，且客户对于投入产出比的敏感性，这个指标也是很容易衡量的。

举个例子，2017 年底联想开始与中国顶尖的石化企业合作，一起做催化裂化装置的工艺参数优化。最后联想的投入是客户出资的五六倍，但是团队也乐于与跟客户一起，把石化行业的标杆客户拿下，这件事对企业有非常大的业务价值，且这类价值比较容易衡量。

今年的疫情期间，虽然全球经济都受到了影响，但是田日辉的团队并没有改变原有的发展路线图和考核标准。

田日辉告诉我们，从用户使用的角度来说，云模式的产品无论从分析还是从报表各方面，基本上没有太大的影响。中央的数据团队虽然在家里远程办公，数据用户也在家里办公，但是整个平台运转还是很正常。



疫情对团队业务本身供应链会有一些影响，即用户的需求。由于疫情影响，客户对联想的产品，无论是业内产品还是云服务的产品，以及数据智能转型服务这些产品的需求，都是有所提升的。联想的私有云给外部客户提供的服务略有影响，体现在团队帮助客户来构建私有云之后，在建设初期，需要做用户访谈和数据治理，和云业务部门有比较多的沟通，这些项目会受一定的影响，但是比例不是很大。在面对疫情等重大社会事件下的经济现状中，田日辉也希望团队保持积极心态，推定数据平台与服务升级，寻求稳定发展。





受访嘉宾：

尚晶

中国移动信息技术中心大数据平台部副总经理

BIG DATA DIGEST 大数据文摘

通信大数据行程卡小程序，相信大家都不陌生。

疫情期间，它不仅是人人手机里必备的小程序，也是外出的必要通行证。

今年 4 月，为了应对企事业单位的大面积复工复产，并且准确掌握居民个人过去 14 天的行程，全国一体化政务服务平台上线工信部推出的“通信大数据行程卡”服务，并将行程卡信息纳入全国一体化平台“防疫健康信息码”服务。用户在信息码服务中申报行程即可查询和证明本人近 14 天的到访地，不再需要另外开具证明。

在工信部统一组织下，三家电信运营商很快实现了数据整合，为疫情期间的出行和公共健康的防疫管理作出了重要贡献，包括中国移动在内的多家国内主要运营商都参与其中。但很少有人知道，整个项目从开启到初版上线其实只花了不到一周。

上个月底，带着对相关团队的好奇，大数据文摘采访了中国移动信息技术中心大数据平台部副总经理尚晶。她所在的中国移动信息技术中心负责中国移动全网 IT 系统统一规划、建设和运营，今年还加挂了中国移动大数据中心的牌子，按“一套人马，两块牌子”运作，目前负责大数据相关工作的团队有近 200 人。

这个团队可以说是中国移动大数据的一支“集中兵力”，不仅负责中国移动集中化大数据平台系统和应用的建设运营和分析支撑，同时还肩负组织各省近 400 多人的大数据团队，推进全网大数据工作的职责。

疫情期间，正是这支数据团队，带领中国移动 31 省大数据精英，与时间赛跑，有效的支撑了疫情人群迁徙、行程查询、复工复产分析等各项工作，累计提供疫情防控分析报表上万张。

而能在短时间内完成这一切，一个高效的大数据平台和数据团队必不可少。

20 年建设经验，数据团队建设“划重点”

中国移动大数据建设还得从大数据系统的前身——经营分析系统开始谈起。

我们从中国移动大数据中心了解到，中国移动的经营分析系统建设从 2002 年开始，技术上采用数据仓库。当时大数据这个概念还没有出现，考虑到初期投资成本较大，国内数据仓库系统的建设主要是电信运营商、银行、保险公司这些百强企业。然而随着数据量爆炸性的增长，一方面 Oracle、Db2 等数据仓库在存储 PB 级数据上开始显现扩展性不足和非结构化数据处理能力不足的问题，另一方面昂贵的价格，也逐渐成为一个亟待解决的问题。投资收益率的问题开始越来越多地被问及，2007 年中国移动研究院首先开始跟进 Hadoop 的研究，2009 年，中国移动开始在省级系统上热火朝天的开始新兴 MPP 技术、Hadoop 技术的试点和大数据平台建设。

这个时期也是互联网公司开始从 IT 时代向 DT 时代演进的前夜。

与运营商的审慎探索不同，互联网企业的成本压力和技术实力，促使他们更快的拥抱了开源体系，例如 2009 年阿里的云梯 1 和云梯 2 项目。

“这不仅仅是一个技术变革和颠覆，背后更是一个生态变革。”

意识到这个问题的中国移动在 2015 年明确了大数据建设的组织机构，大规模推进集中化大数据平台的建设，并推进自研 BC-Hadoop 在现网的落地应用，单集群规模迅速从 300 台扩展到 3000 台，整体规模达到 2.5 万节点，集团大数据平台的采集数据量从 2015 年的 20TB/ 日，到达 1.9PB/ 日。应用领域上，更是从决策支撑 + 营销支撑为主，不断向外拓展，内部深入到企业运营的各个领域，包括精细服务、产品创新和高效运维等，向外拓展金融、旅游、交通、零售、安全等多个垂直领域合作。

尽管系统建设速度和应用构建速度较之之前近乎按数据级提升，但是“还需要大数据支持”的声音仍然在中国移动的各个层级的单位机构中此起彼伏，2017 年开始，集中化大数据平台开始小规模的推广大数据 PaaS 开放模式。这个开放平台被命名为“梧桐”平台，意在“梧桐花开，凤凰自来”，提供储算资源、大数据处理工具、全网汇聚数据和安全管理能力，向内部各单位开放赋能。

“梧桐”平台一经推广，得到了巨大的响应，短短一年内就从几个单位入驻，迅速实现了 50 个覆盖省公司、专业公司的数百个项目的入驻。而与此同时变化的是，这个数据支持团队，也悄悄的拆成了大数据平台部和大数据应用部两个部门，两个部门均在近百人团队，以适应更为开放的服务生态。

从“授之以鱼”的应用提供方式，到“授之以鱼”+“授之以渔”结合的方式，需求部门可以自行选择“买鱼”还是“买船自己出海打渔”。

在中国移动大数据中心，这些变化也是在对大数据工作各种困惑的思考中，不断摸索优化推进的，很多改进都是大数据中心领导亲抓亲管。所有的努力从更理论一点的角度看，其实都是在思考如何将国家新明确提出的“数据要素”真正做到要素化，让数据能安全的流动起来，流动到所需要它的地方，流转到企业内部每个需要数据赋能的环节，也包括数据要素在跨行业合作中产生“化学反应”，创造新的产品、新的价值。

目前梧桐已经成为中国移动数据中台的品牌，引入更多的新技术提升数据中台的计算效率和实时性，提高数据中台开放敏捷性、易用性，加速应用创新是目前团队考虑的主要问题。



平台搭建好之后，数据团队需要进一步考虑的就是如何衡量大数据的价值。尚晶表示，这个问题是从经分时期就一直在被问的问题，但或许到现在也仍没有一个完美的答案，目前主要有以下几个考量角度：

1. 渗透行业领域的广度考核，比如金融行业、零售行业、交通行业、旅游行业、公共安全行业等行业，形成了哪些赋能应用。

2. 带来的经济价值或者间接经济价值，例如由于采用大数据，同等营销资源投入下，营销成功率的提升，大数据分析发现的收入漏损，大数据直接产生的政企行业合作收入。较难计算的是间接经济价值和拓展行业的机会成本，例如企业专线销售与打包的大数据服务，收入占比较难衡量。又如基于大数据分析，面向市场设计的产品，多少价值应该计入大数据带来。

尚晶也给我们举了个例子。普通的营销方案成功率可能就在 1%，在流量市场这个比例甚至更低，但无论营销成功与否，营销成本还是需要花费的，比如外呼人员成本、短信端口信息成本、优惠券成本，这些都是成本消耗。如果采用大数据分析，可以得到一些更精准的目标群体，根据这些有针对性的有效目标群体做营销，成功率就会从 1% 上升到 5%，提升了 5 倍，同等成本获取了更多的营收。公司给的营销费用要和成本费用一样，需要和收入一起纳入考核。

辅助业务部门决策，分析师要懂业务更要懂用户

大数据可以发挥价值的角度多种多样，这一点毋庸置疑。但聚焦到辅助业务部门做“数据驱动”的决策这一工作上，中国移动也探索了自己一套行之有效的运作方式。

我们从中国移动大数据中心了解到，中国移动大数据中心有一个分析师团队，他们除了为市场等业务部门提供各类分析数据，还会基于数据

去深度挖掘业务中存在的问题。这个分析师团队目前大概有 20 多人，每周为公司领导提供覆盖全网、不同角度、不同领域的分析，这种分析有效对公司高层的决策起到很好的参考支持作用。

例如在市场竞争中，中国移动部分省公司的客户流失率或价值流失影响很大，那么数据团队就会去分析，为什么流失率这么大，省间差异的原因？移动能与其对标的产品套餐是怎样的？主要流失用户的特征是怎样的？如何发现客户在离开之前的行为异动，及时沟通挽留？又如中国移动咪咕阅读的推广，如何将用户分类，青少年、中年人等人群的阅读喜好，如何分析竞品业务数据？如何引入更好的内容，并精准推荐，保持用户粘性。

除了集中的分析团队，中国移动在各省和专业公司内部也在推进业务与大数据分析的融合团队，发挥整体优势，面向实际业务运营，充分发挥大数据价值。例如中国移动向用户提供 2018 央视世界杯新媒体直播权益推广，数据团队就会将世界杯比赛时间和球队粉丝的活跃度进行关联，并挖掘其中的必然联系，用其中的联系特征来做营销方案，并根据用户人群进行有效划分和推送。比如是青少年，那么应该推送应援物资的售卖渠道，如果是中年人，那么应该注重的是中年人更为关心的内容，进球精彩瞬间等。又如 5G 营销，各省也是先进行了客群定位，对客群和潜在市场进行分析，再制定营销计划，包括做营销的排期，营销的资源投入，营销渠道的资源顾客。推送之后，可能会产生沉默用户，这个时候再做沉默分析，比如分析出时间不对，一边采用大数据分析结果一边调整。

“做分析必须要懂用户心理。”

比如说中国移动的花卡推广，面向的是热爱娱乐的青年群体，分析师需要从青年群体的喜好角度去分析，才能充分获知用户购买动机，更倾向的优惠促销品，洞察业务设计中合理性。

在懂得用户心理与需求的情况下，分析师需要更为多元以及完整的数据。分析师除了要做分析以外，还需要对业务深入理解，分析师会需要一些来自数据团队的支撑，比如对业务数据的解释以及根据数据得到的建议。在中国移动除了分析师团队，还在打造数据团队，两个团队以数据需求为纽带形成持续的数据应用与探索 + 新数据引入与治理分工协作的良性循环。

用户行程分析，数据安全如何保证？

除了分析用户的喜好和日常来分析业务，疫情期间，为了公共安全和安全防疫，中国移动的数据团队也全程参与了通信大数据行程卡项目。通信大数据行程卡是基于用户位置数据的，因为数据相对敏感，在技术保障和用户授权问题上，中国移动的数据团队也时刻把用户的隐私数据放在第一位。

中国移动大数据中心的处理方式有以下几个原则，首先根据网络安全法，采集数据使用数据必须都得到用户的授权。用户的授权体现为用户入网的时候的协议和合约，移动为用户提供服务时会采集一些数据。在使用用户数据时，会再次请求授权，并明确告知数据用途，比如像采集用户对内容的喜好以及相关的数据，根据这些数据对用户做一些推荐，如果没有用户的授权将无法运用数据。用户可能已经留意到行程卡等用户数据查询，都通过短信确认码或要求输入身份证后 4 位，作为用户二次确认依据。

除了在用户授权和安保措施以外，中国移动数据团队对数据安全也有做非常多的技术方面的措施。数据在系统里均为加密存储，并按需进行了模糊化和脱敏处理，数据访问权限按最小授权原则，数据操作遵守严格的安全审核审计金库管理模式。数据分析人员无法了解数据与真实用户的关联，因此可以保证对个人客户隐私数据的充分保护。

给好的数据团队下个定义？

采访的最后，我们也请尚晶给“好的数据团队”下个定义。尚晶告诉我们，其实她一直在思考这个问题，回答好这个问题才能明确团队未来努力方向。

但这不是一个容易回答的问题，需要放到快速变化的、公司内外部、技术与生态的环境里去思考。一个团队成功要有别人难以超越的长板，但是一个团队的长久成功需要没有明显的短板。

“对于成功的数据团队，有很多取得共识的分享，包括从组织上、管理机制上、技术水平上、数据能力、应用价值、行业口碑、市场收入上，数据中台的争论已经有各种反转又反转。归根结底还是成功的中台经验是相近的，而失败的中台各有各的失败，也就是短板”。回归到 IT 本职和她所从事的大数据中台工作，尚晶希望从三个层级去描述大数据的评价体系：“数据融合”、“开放共享”、“赋能创新”。

数据融合：数据覆盖范围是否充分？是否建立完善的数据管理体系，有

效保证数据完整性、可靠性和及时性？是否有先进的技术架构，有效捕捉业务数据，实现高效储算并敏捷为业务提供数据服务调用？

开放共享：是否适应复杂的需求场景？是否有适配公司的组织机构的开放模式？是否有高效的复用度和支持度？是否有开放的数据字典，可为使用人员充分理解？是否有丰富、便捷部署，易用性好的工具？是否有敏捷的、有 SLA 保障的开放流程？

赋能创新：是否能有效赋能公司的目标市场，就中国移动而言即 CHBN 四轮市场？是否有助于公司创造新的增长点？是否彰显国企服务民生的担当？

这也是中国移动大数据中心对数据团队未来发展方向的期望。尚晶认为数据团队首先还是要配合业务的发展，需要对行业进行深入挖掘和分析，还有对客户群体的深入分析，满足客户不断增长的新需求。

“中国移动的大数据，发展潜力还很大，还有很多值得探索，做深做广的领域，未来中国移动集团公司也要求在数据团队加大人才培养力度，建立更加灵活的机制选聘行业专家加盟，共创未来。”

为日均服务十亿人次做准备，美团数据团队如何走在业务前想问题



受访嘉宾：

李闻

美团数据平台负责人

“你每一次花钱，都是在为自己想要的生活投票。”

2010 年 3 月 4 日，美团网站上线当天，美团创始人兼 CEO 王兴发出了这样一条微博，希望以“吃”为核心，去打造一个帮大家吃得更好，生活更好的全方位生活服务平台。

当然，要协调日订单已经突破 3000 万单的外卖配送以及包括快驴、买菜、单车、酒旅在内的多个业务线，一个稳定、强大的数据基础架构必不可少。

王兴给美团定的下一个目标是每天服务十亿人次，这个并发量对美团数据团队来说，将是不小的挑战。但同时，为了应对异常复杂的业务场景，保证跟技术的极致融合，美团数据团队也发展出了自己独有的特点。

“指挥部”核心支撑，“小兵团”灵活作战

据美团数据平台负责人李闻介绍，从宏观方面来说，美团内部整个大数据团队主要涵盖两大技术方向：一个是数据研发方向，涵盖面向数据资产的数据清洗、加工、整合、挖掘、管理、运营等技术领域，主要包括批处理和实时数据仓库的建设、数据管理、数据价值落地以及数据运营；另一个是数据系统研发方向，涵盖批处理、实时数仓开发工具链、BI 系统、数据管理系统等数据系统研发。

大数据团队作为一个整体，希望通过数据内容建设、数据系统建设，来提升美团整个公司数据质量、数据效率、数据安全，以数据驱动的方式帮助公司完成业务目标，持续提高公司的运营效率和核心竞争力。

但是，涉及到实际的业务时，美团跟不少单一业务线公司“数据团队中央化管理”又有所不同，他们采取了基础研发部以“指挥部”的形式核心支撑，各业务线通过自有的规模较小的嵌入式数据闭环形成“小兵团”，灵活高效的完成单线任务。

李闻说，美团是多业务线多 BG 的组织形式，每一个业务线都有自己的研发团队，即每个 BG 下面有自己的数据工程师（DE）和数据科学家（DS）。其中工程师团队主要负责中心化的公共数据建设，而数据科学家团队则是面向公司集团层面的经营分析和决策，一些涉及公司重大发展方向的战略性问题，都会优先进行数据分析再进行决策。

而在每个业务的“小兵团”之下，也有一个中心化的大数据团队，服务对象是全公司所有的业务线，为全公司所有的 BG 业务线提供能力支撑，这点与其他互联网公司相比也有很大的不同。

这个中心化的大数据团队，对全公司所有业务线提供全公司统一的数据技术平台和公共数据内容平台支撑，以及面向集团的商业分析支持，除此之外，美团中小业务在孵化阶段，也由这个大团队提供人力、技术资源支撑，快速建立数据能力。

这一组织形式是由美团复杂的业务场景特点决定的。

美团目前有超过 200 个生活服务场景，每个场景都具有自身的业务特点和数据维度特点，如果只是通过平台式的数据中心来进行相关处理，无法实现最高效的数据处理和灵活的技术迭代。而中心的平台能够在其中实现最大限度的资源协调，并从集团层面处理可复用性的公共数据，整体负责整个公司的公共流量、公共维度，还有一些和用户相关的用户画像都数据内容。

以美团金融服务业务为例，数据工程师的工作职责包括以下几个方面：

- 搭建并优化金融服务数据体系，包括数据仓库、数据应用和实时统计等系统的开发，及对安全性、存储计算成本、查询性能和使用体验进行综合优化。
- 参与商业智能系统建设，建设 PB 级高效、灵活的在线分析、自动归因和智能预测。
- 为各类业务场景提供综合数据解决方案，包括数据生产采集、安全合规、实验设计、评价监控、数据挖掘和智能决策等。

对美团来说，金融服务是极重度的数据型业务，业务的高效运行和有效决策都依赖于数据技术的支撑。另一方面，数据是金融科技的前沿，美团希望通过互联网数据技术的发展和应用，帮助合作的金融机构提升技术生产力，从而促进整个生态的发展。

四大发展阶段，数据团队承担着不同的角色

作为一家非常重业务的公司，美团业务经营核心诉求包括战略决策、经营策略、运营策略（从人工运营到机器运营），而这些都离不开数据的支撑。

但是，随着信息技术的发展和普及，产生数据的信息源越来越多，获得洞察所需要的信息也越加丰富，但是这些错综复杂甚至是无序不规范设计的信息系统的数据是不一致的、分散的，所以就需要有一个非常重要角色把这些数据进行重新的清洗、整合，形成统一商业视角下的数据“模型”。

访谈过程中，李闻也从“互联网业务”整个生命周期的视角解读了数据和的价值和数据团队在这一周期过程中的发展阶段：

1. 初创期：这是业务从无到有的阶段。此时企业经营的重点是找到让人信服的商业模式。对研发的诉求主要是后台和前端，让面向用户的产品能够运转起来。此时公司对数据的诉求主要是一些基础指标的表现，用以判断商业模式的合理性，往往让了解数据产生机制的后台和前端同学承担数据统计工作就可以了。当然，在基础比较好的团队里，可以通过敏捷的统计工具直接连接数据源，写 SQL 统计数据并做基本的数据展现。基础类的数据指标工具比如美团的“魔数”在此时发挥的作用最大，属于一个基础设施。

2. 成长期：在这个阶段，商业模式已经被证明是可行的，进入扩张规模抢占市场阶段。业务规模快速膨胀，此时的数据量也随之大量增加，需求也在不断迭代。既要保障现有任务的稳定性，还要快速支持蜂拥而至的需求，需要打好数据基础，做好需求管理。该阶段是对数据技术压力最大的阶段，更多是如何高效应对需求且保障现有任务的稳定性和数据的准确性。

3. 成熟期：在保障规模下追求“毛利”为正。此时，企业经营的基本思路已经成型，需要系统建设指标体系，利用数据科学严谨的指导经营，并利用用户画像等技术更精细地理解用户从而精准营销，提高运营 ROI。此时需要做好数据的治理以及内容的体系化管理。比如美团数据中台就是在这个阶段演化出来的。

4. 持续发展期：这属于通过数据来扩大利润的阶段，企业需要结合对业务的深刻理解和行业的发展趋势，采集和整合更多元的数据内容，结合本业务特点，发现高价值用户、挖掘更多商业机会、输出更多增值服务，丰富业务的利润结构。此时，还需要更深刻的理解用户，理解数据，通过数据产生更多洞察，提高经营效率。数据开发领域的终极发展目标，应该是懂数据开发（集成）技术，懂产品的业务逻辑，懂商业，懂分析，懂经营策略，懂运营策略，同时还能推动各相关角色配合行动的综合性人才。数据源越复杂，为保障交付数据的准确性，挑战就越大，数据开发的核心价值就越大。

四大发展阶段，数据团队承担着不同的角色

“走在业务线前面主动去做一些工作，每当业务碰到的问题时，最好平台都有解。”

在谈到如何定义一个好的数据团队时，李闻如此回应。因为美团属于跨业务线、多 BG 的模式，这让每个业务线的数据都存在很大的可复用性。那么，如何在兼顾安全的前提下，让各业务线能够更高效地用到跨团队数据呢？这也是业界不少数据团队在建设初期面临的一个比较棘手的问题。

美团目前的解决方案称之为“分场景分角色安全域”，即在整个数据体系中按照数据、算法、商业分析分场景分角色建立安全域，在保障数据安全的同时，简化授权模型，建立起一套比较清晰的数据权责体系，减少数据供给方和需求方的数据交换成本。

另外，沟通机制和认知提升也很重要，美团数据团队不倾向于把大数据和业务线分隔得太清晰，一方面直接深入到业务线，积极响应每一条业务线的需求，另一方面也在构建底层基础能力，大力研发，不断进阶，为未来的业务发展做好充足的准备。

大数据平台是重要“基础设施”，支撑 AI 和大数据两条线

此外，为了让整个工程团队和基础架构团队能够最大效率地发挥效用，美团的大数据平台和机器学习平台是在组织和平台技术上是重合的，这种设置在业界也非常少见。

众所周知，AI 是目前互联网领域炙手可热的“明星”，无论是老牌“巨头”，还是流量“新贵”，都在大力研发 AI 技术，为自家的业务赋能。

在刚刚过去的世界人工智能大会上，美团首席科学家夏华夏首次公开呈现了美团 AI 的建设图谱，在这一图谱中能看到李闻所在的大数据平台部门是美团 AI 建设的一个重要“基础设施”，同时支持着大数据和机

器学习两条线。他认为，公司数据团队之所以发展成这种形式，从本质上讲，是因为大数据和机器学习两个领域底层的基础设施和能力实际上是可以“共用”的，包括一些工程方法也比较类似。

李闻说：“大数据和机器学习平台技术，其实在技术角度没有清晰边界，在其他公司强行拆在两个团队，更多是组织和人的原因。”



“比如做数据清洗，一样会用到数据挖掘算法，做一些深度学习中前置的特征处理或者特征准备，实际都在用大数据的技术。其次，大数据和机器学习底层的一些架构技术、工程方法和能力模型实际是很类似的，包括一些分布式的技术，都是可以复用的。美团这种组织形式，在实际工作中，确实对提高工作效率有非常大的助力。”

数据治理老大难，在支撑和治理间寻求平衡

由于美团的业务线众多，应用场景也非常复杂，跟其他互联网公司一样，美团也在同样面临着数据治理的问题。在业界，数据治理有两大难题：数据资产治理和数据成本治理，其中数据成本相较于业务成本的投入会呈现长期累加的特征。那么，如何在效率和成本之间找到平衡，李闻详细讲述了美团采取的自主摸索的方法。

据李闻介绍，从数据源头整个加工到产出报表再到使用，其链条会非常长，涉及的角色也非常多，变量也很多，伴随着业务系统的变化，中间的数据逻辑，以及指标口径定义也会随之变化。如何去管理这样的一些变化，去拿到一些预期的数据结果，就是一件非常具有挑战的事情。

数据平台团队作为公司中心化技术团队，同时需要扮演两种角色，一方面要以客户为中心，提供能力支撑好公司各业务在大数据和算法领域的工程技术需求，另一方面同时要扮演公司的治理抓手，驱动整个数据、算法体系成本、效率、质量、安全的提高，“我们本质上有两拨客户，一波是公司各业务数据、算法研发，一波是公司管理层，同时满足好两拨客户的诉求，是需要极大的韧性、极强的技术能力的”，在谈及数据治理问题时，李闻表示。

资源内部按钱结算是美团在成本治理层面所使用的核心策略。在这一策略的支撑下美团在 2017 年就已经做到了内部的云化和资源按钱结算，在美团内部，数据平台对每种资源类型都会有定价，各条业务线技术负责人提出储存和计算的需求，业务线 BM 可以直接看到本业务线在大数据上花多少钱，数据平台会从技术视角 Review 资源需求的合理性，最后结合全局优化目标将资源转化为机器采购，提交给云计算。数据平台除了作为公司大数据成本的技术把关人，同时也提供能力和工具支持各业务线成本优化，以及在底层引擎层面做持续的迭代和优化，底层引擎每年都会有接近 10% 的效率提升。

“实际上，通过这样一套机制能持续去推动每一个业务线去做优化。因为每一个业务线都有一套自己的商业模型，要去核算自己的成本和收益，你只告诉他们花多少资源，花多少机器，实际上是没有什么帮助的。”

如果从这一角度而言，在业界，美团算是一个“先行者”。

从支撑业务到驱动业务

目前，美团的数据平台技术体系，早已经度过了“基于开源搭一搭，魔改一下就能解决问题”的阶段，业界开源技术已经不能满足业务需求，需要在部分领域构建能力做自研。另一方面，也度过了“对外对标业界技术、学一学就能坐时光机少走弯路”的阶段，由于美团业务特点和发展阶段，数据平台技术领域碰到的问题，很多是独特的，通过对标业界已经无法获得更多有效输入，已经需要通过紧密结合业务问题和领域技术发展趋势，向内深度自我洞察、自我反思，在领域内自我技术突破、从工程技术支撑业务到工程技术驱动业务的转变。

“美团整个数据平台技术在业界应该还是比较靠前的，例如整个架构

技术，很早就解决了大规模数据复杂关联场景多地域的平滑扩展性问题。”在谈及美团技术优势时，李闻表示，“我们很早就做完了计算引擎的内存化升级，持续做计算效率的一些迭代，在成本治理领域是比较独特的。另外，在整个工具层面实际是一套平台，一个大的集群。而其他很多公司只是一些小的平台或者小的自建的集群，数据打通共享是个大问题，当然这跟公司的发展阶段有很大的关系。在机器学习训练部分,我们可以做到700并发0.7倍的加速比，推理部分 BERT 模型性能可以超越业界 state-of-the-art 1～2 倍的样子，虽然取得了一些成绩，但是未来的挑战也很大，美团数据团队还是会本着求真、务实的心态，长期有耐心，去迎接这些挑战。”



互联网下半场，数据团队的未来

2020 年 3 月 4 日，美团迎来十周岁的生日。



根据美团 2019 年年度财报显示，美团平台上有单骑手数量已经达到了 399 万，高峰期外卖日订单量达到了 3000 万单，超过 4.5 亿的用户在美团上获取生活服务，而线上有超过 610 万的商户……这些数据背后能带来的产出对美团来说是一笔重要的财富。

也正是因此，数据团队在美团的位置举足轻重。问及在权衡数据团队的商业影响力方面的思考，李闻提到数据技术团队的 KPI 主要看两部分，一是能不能支撑好全公司所有数据团队的工作，比如开发效率、数据使用效率等；二是要考虑与全局数据成本、全局数据质量相关的一些 KPI。



李闻说：“在美团有一条非常重要的价值观，就是追求卓越。未来的路还很长，美团数据团队也希望能够挖掘出更多的数据价值，并将这些价值转为生产力，帮助公司乃至帮助社会提升效率，创造出更大的价值。”

数据来源

领英、猎聘、拉勾、智联招聘、前程无忧、中华英才网、阿里巴巴官网、腾讯官网、百度官网、京东官网、美团官网

数据统计时间：2020 年最新数据



版权声明

本报告版权和最终解释权归清华大学大数据研究中心及大数据文摘所有，职位数据获取自领英平台。未经许可的转载以及改编者，我们将依法追究其法律责任。

引用、转载、合作，请联系 zz@bigdatadigest.cn，或电话 010-86463811

报告制作团队

主编团队
魏子敏 赵玮雯 刘俊寰 牛婉杨

深度访谈团队
魏子敏 刘俊寰 牛婉杨 夏雅薇 刘纯 耿冉 朱玲

问卷调研团队
曹培信

数据分析及可视化团队
赵玮雯 邓耀

致谢

赖春波	滴滴技术副总裁、数据科学与智能部负责人
许 亚	领英全球数据科学团队负责人
黎 巍	腾讯安全副总裁
田日辉	联想数据智能业务集团产品及生态总经理
尚 晶	中国移动信息技术中心大数据平台部副总经理
李 闻	美团数据平台负责人
汪德诚	大数据文摘创始人
寒小阳	互联网资深算法专家




BIG DATA DIGEST
大数据文摘
清华大学软件学院


BIG DATA DIGEST
大数据文摘
清华大学软件学院


BIG DATA DIGEST
大数据文摘
清华大学软件学院





考研资料：数学、英语、政治、管综、西综、法硕等（整合各大机构）
英语类：四六级万词班专四专八雅思等
财经类：初级会计、中级会计、注册会计、高级会计、税务师、会计实操、证券从业、基金从业、资产评估、初级审
公务员：国考、省考、事业单位、军队文职、三支一扶微信 2270291360
银行：银行招聘、笔试、面试
教师资格：小学、中学、教师招聘面试
建筑：一建、二建、消防、造价
法考：主观题、客观题
多平台网课：涵盖职场、办公技能、编程、文案写作、情感心理、穿搭技巧、理财投资健身减肥摄影技术等优质内容
精选资料：Excel 教程、PPT 模板、简历模板、PS 教程、PPT 教程、素描、烹饪、小语种、CAD 教程、PR 教程、UI 课程、自媒体、写作、计算机二级、钢琴、Python、书法、吉他、kindle 电子书、演讲.....持续更新中...
押题：提供考前冲刺押题（初级会计、中级会计、注册会计、一建、二建、教资、四六级、证券、基金、期货等等），麻麻再也不用担心我考不过了。
资料领取微信：1131084518
行业报告：20000 份+持续更新

英语四六级备考资料	计算机二级备考资料	150 所高校考研专业课资料
两小时搞定毛概马原思修近代史纲	高数(微积分)+线性代数+概率论	素描 0 基础入门教程
教师资格证全套备考资料	普通话考试资料礼包	书法教程微信 2270291360
大学生英语竞赛备考资料	大学生数学竞赛备考资料	1000 份各行业营销策划方案合集
挑战杯/创青春/互联网+竞赛资料	电子设计竞赛必备资料	街舞 0 基础入门教程
托福雅思备考资料	大学物理学科攻略合集	动漫自学教程
SCI 最全写作攻略	TEM4/TEM8 专四专八备考资料	教师资格证面试试讲万能模板
360 份精美简历模板	数学建模 0 基础从入门到精通	100 套快闪 PPT 模板
Vlog 制作最全攻略	超强 PR 模板	42 套卡通风 PPT 模板
PS 零基础教程微信 1131084518	PS 高级技能教程	63 套酷炫科技 PPT 模板
好用到极致的 PPT 素材	128 套中国风 PPT 模板	32 套 MBE 风格 PPT 模板
327 套水彩风 PPT 模板	295 套手绘风 PPT 模板	54 套毕业答辩专属 PPT
196 套日系和风 PPT 模板	82 套文艺清新 PPT 模板	57 套思维导图 PPT 模板
163 套学术答辩 PPT 模板	53 套北欧风 PPT 模板	34 套温暖治愈系 PPT 模板
118 套国潮风 PPT 模板	30 套仙系古风 PPT	126 套黑板风 PPT 模板
114 套星空风格 PPT 模板	192 套欧美商务风 PPT 模板	42 套绚丽晕染风 PPT
50 套精美 INS 风 PPT 模板	56 套水墨风 PPT 模板	137 套清爽夏日风 PPT 模板
98 套森系 PPT 模板	25 套简约通用 PPT 模板	记忆力训练教程
300 套教学说课 PPT 模板	123 套医学护理 PPT 模板	AE 动态模板微信 2270291360
毕业论文资料礼包	教师资格证重点笔记+易错题集	表情包制作教程
吉他自学教程（送 6000 谱）	钢琴自学教程（送 1000 谱）	区块链从入门到精通资料
2000 部 TED 演讲视频合集	Excel 从入门到精通自学教程	单片机教程
230 套可视化 Excel 模板	1000 款 PR 预设+音效	1000 份实习报告模板
手绘自学教程微信 1131084518	单反从入门到精通教程	人力资源管理师备考资料
英语口语自学攻略	粤语 0 基础从入门到精通教程	证券从业资格证备考资料
日语自学教程	韩语自学教程	PHP 从入门到精通教程
法语学习资料	西班牙语学习资料	炒股+投资理财从入门到精通教程
全国翻译专业资格考试备考资料	BEC 初级+中级+高级全套备考资料	大数据学习资料
SPSS 自学必备教程	Origin 自学必备教程	会计实操资料
LaTeX 全套教程+模板	EndNote 教程+模板	小提琴 0 基础入门自学教程
GRE 超全备考资料	200 份医学习题合集	司考备考资料

思维导图教程＋素材	20000 份行业研究报告	一级造价工程师备考资料

上万 GB 教学资料（均全套,非杂乱）免费领取微信 1131084518

《闪电式百万富翁》实战版+升级版

易经+道德经+易学名师全集+风水学+算命学+起名+++等等（全套 1000 多 GB）

心理学+NLP 教练技术+精神分析+亲子家庭教育+催眠+++等等（更新超 2000GB）

大学-已更新至 9333 个课程+高中+初中+小学-全套资料（超过 2 万 GB）

陈安之	曾仕强	马云	杜云生	翟鸿燊	刘一秒	俞凌雄
王健林	余世维	雷军	周文强	安东尼罗宾	董明珠	李嘉诚
徐鹤宁	冯晓强	李践	刘克亚	罗伯特清崎	戴志强	李伟贤
苏引华	史玉柱	李强	俞敏洪	杰亚伯拉罕	周鸿祎	唐骏
梁凯恩	陈永亮	傅佩荣	贾长松	易发久	李彦宏	湖畔大学
李开复	聂枭	悟空道场	魏星	姬剑晶	其他名师全集	其他资料下载
王兴	王智华	智多星	陈文强	周导		微信 2270291360
泡妞	撩汉	泡仔	房中术	性福课	驻容术	泡妞撩汉性福合集

注：太多了，无法全部一一列出。。。

全套专题系列【微信 1131084518】

记忆力训练	形象礼仪	健康养生	企业管理	沟通技巧
演讲与口才	经理修炼	MBA 讲座	时间管理	战略经营
企业文化	销售心理	管理素质	国学讲座	执行力
团队管理	领导艺术	员工激励	潜能激发	谈判技巧
绩效管理	薪酬管理	43 份直销制度	电话销售	人力管理
客户服务	创业指南	市场营销	餐饮管理	保险讲座
品牌营销	酒店管理	汽车 4S 店	众筹资料	销售技巧

兴趣爱好：	钓鱼教程	魔术教学	炒股教学	美术教学	书法教学	
音乐乐器：	萨克斯教学	电子琴教学	小提琴	古筝教学	钢琴教学	
	吉他教学					
体育运动：	篮球教学	足球教学	羽毛球教学	乒乓球教学	太极拳教学	
	围棋教学	高尔夫球				
生活实用：	插花教学	茶艺-茶道	唱歌教学	单反相机摄影	毛线编织	
	小吃+美食					暗号：666
语言学习：	英语					免费领取资料微信
电脑 IT：	办公 office	PS 美工教学				1131084518

微信1131084518
撩汉liaohan.net
最好资源zuihaoziyuan.com





START YOUR FINANCE



起点财经，网罗天下报告